وزارت ارتباطات و فناوری اطلاعات
پژوهشگاه ارتباطات و فناوری اطلاعات

راهکارهای گردآوری داده و ارائه تحلیل‌های توصیفی از ترافیک شبکه های WAN

تهیه‌کنندگان: مریم امیری، ناهید امانی

عنوان گزارش: راهکارهای گردآوری داده و ارائه تحلیل‌های توصیفی از ترافیک شبکه‌های WAN

تهیه کننده: مریم امیری، ناهید امانی

گروه پژوهشی: گروه مدیریت یکپارچه شبکه

تاریخ نشر: ۱۴۰۲

فهرست مطالب

۱مقدمه ۴

۲ داده‌های مورد نیاز برای پایش ترافیک شبکه ۶

۲-۱ لایه‌های شبکه و انتقال داده‌ها ۶

۲-۲ جمع‌آوری داده‌های ترافیک شبکه ۷

۲-۲-۱ جمع‌آوری لاگ از نودهای سوئیچ و روتر شبکه ۸

۲-۲-۲ جمع‌آوری لاگ از سیستم‌های مدیریت شبکه ۹

۲-۲-۲-۱ پروتکل Syslog ۹

۲-۲-۲-۲پروتکل SNMP ۱۰

۲-۲-۲-۳پروتکل NetFlow ۱۰

۲-۲-۳جمع‌آوری داده‌های ترافیک وب ۱۲

۲-۲-۳-۱پروتکل‌های اصلی وب(HTTP و HTTPS) ۱۲

۲-۲-۳-۲جمع‌آوری لاگ سرورهای میزبانی مراکز داده ۲۰

۲-۲-۳-۳جمع‌آوری لاگ ربات‌ها و اسکریپت‌ها ۲۰

۲-۲-۳-۴جمع‌آوری لاگ از فایروال‌ها و سیستم‌های امنیت شبکه ۲۲

۲-۲-۳-۵جمع‌آوری لاگ از سرورهای DNS ۲۴

۲-۲-۴نمونه‌گیری از داده‌های ترافیک شبکه ۳۱

۳تحلیل ترافیک شبکه ۳۳

۳-۱پایش و مدیریت شبکه با ارائه تحلیل‌های توصیفی بر روی لاگ ترافیک شبکه ۳۴

۳-۱-۱تحلیل دسته بندی محتوای متنی وب‌سایت‌های بازدید شده توسط کاربران شبکه ۳۵

۳-۱-۱-۱الگوریتم‌های یادگیری ماشین بانظارت با استفاده از خزش محتوای آدرس‌های ULR مورد بازدید کاربران ۳۶

۳-۱-۱-۲دسته‌بندی مبتنی بر قوانین ۳۸

۳-۱-۲تحلیل‌های آماری روی ترافیک بازدید کاربران از دامنه‌های وب ۳۹

۳-۱-۲-۱معیارهای تحلیل آماری ترافیک وب ۳۹

۳-۱-۲-۲نمونه‌هایی از تحلیل آماری بازدید کاربران شبکه از دامنه‌های وب ۴۱

۳-۱-۲-۳نمونه‌هایی از تحلیل آماری روی حجم ترافیک تبادلی در شبکه ۴۲

۳-۱-۲-۴تحلیل‌های آماری با استفاده از منابع داده‌ای تکمیلی ۴۳

۴جمع‌بندی ۴۴

۱ مقدمه

در قرن حاضر، فناوری مبتنی بر توسعه‌ی عملی، زمینه‌ی اقدامات پژوهشی برای دانشگاه‌ها بوده و به نحوی گسترده راه‌کارهای عملی خود را در حوزه‌های تخصصی جستجو کرده و این امر به مرور جامعه و ساختارهای سازمانی و صنایع را تحت تأثیر قرار داده است. از طرفی در عصر فناوری، داده‌ها و فناوری اطلاعات، شبکه‌های کامپیوتری نقشی مهمی را در توسعه عملی فناوری‌ها ایفا می‌کنند. شبکه اینترنت یا شبکه‌های اینترانت بزرگ مانند شبکه ملی اطلاعات به عنوان شبکه شبکه‌ها، آخرین برگ ارتباطات در آستانه قرن بیست و یکم بوده است.

شبکه های ارتباطی نوین از نظر طراحی، پیاده سازی و ظرفیت به سطحی از پیچیدگی رسیده‌اند که نیاز به مدیریت و پایش پیوسته جهت رسیدن به حداکثر میزان بازدهی را دارند. در طی چند دهه گذشته پیشرفت‌های قابل توجهی در فناوری جمع آوری داده‌ها، تحلیل آماری، محاسبات ابری، یادگیری ماشین و داده کاوی جهت تحلیل‌های هوش مصنوعی داده‌های شبکه، صورت گرفته است. فرآیند جمع‌آوری داده از شبکه‌های ارتباطی و ارائه انواع تحلیل‌های آماری و هوشمند در مقیاس‌های مختلف برای انواع شبکه‌ها از یک شبکه خانگی یا محلی گرفته تا شبکه‌های ملی و جهانی قابل انجام است و می‌توان گفت راه‌کارهای مشابه و تعمیم‌پذیری دارد.

یکی از مهم‌ترین کاربردهای تحلیل ترافیک شبکه‌ها، مدیریت و پایش شبکه‌ها است که بدین منظور داده‌های ترافیکی مرتبط با آن بایستی گردآوری گردد. این داده‌ها اغلب از لایه سوم تا هفتم شبکه‌ها(طبق مدل مرجع OSI) در ترافیک شبکه موجود هستند و کامل‌ترین اطلاعات پایش را می‌توان از لایه کاربرد که شامل داده‌های کاربران است، استخراج نمود. در واقع با استقرار نرم افزار جمع‌آوری، ذخیره‌سازی، پردازش و تحلیل داده مرتبط با همه NMSهای موجود در شبکه و با بهره‌گیری از الگوریتم‌های هوش مصنوعی می‌توان پایش و مانیتورینگ جامعی را بر کل شبکه اعمال نمود.

در شبکه ملی اطلاعات، پایش شبکه‌ها مبتنی بر تحلیل آماری و هوشمند ترافیک شبکه‌ها در اقدامات و اهداف زیر که در اسناد طرح معماری و اصول حاکم بر طراحی و تبیین الزامات شبکه ملی اطلاعات آورده شده است، کاربرد دارد:

- رصد و پایش شبکه ملی اطلاعات
- توسعه ابزارهای تحلیل جریان ترافیک در سطح اپراتورهای شبکه
- بهینه سازی توزیع ترافیک در شبکه و تمرکز ردیابی (داخلی و بین المللی)
- توسعه سامانه رصد، پایش، تشخیص و تحلیل و مقابله با برخی تهدیدات و حوادث شبکه مانند مدیریت خرابی ها
- ابزار تحلیل بازدیدکنندگان و تبلیغات وب برای کسب و کارهای داخلی و خارجی لایه های مختلف شبکه ملی اطلاعات
- اعمال مدیریت یکپارچه شبکه در مدیریت منابع و تخصیص و بهره برداری بهینه از منابع ملی و نظارت بر آن
- تضمین کیفیت از طریق فراهم آوری امکان مدیریت انتها به انتها در شبکه و ایجاد امکان توسعه رقابت مبتنی بر کیفیت خدمات در بین ارائه دهندگان مختلف

- رصد کمی و کیفی خدمات
- ایجاد قابلیت جمع آوری، پردازش و تحلیل داده های عظیم در مورد انواع خدمات و محتوای شبکه و کاربردهای مختلف
- مدیریت ارتباط با مشتریان

به عنوان مثال برای تحلیل جریان ترافیک، بهینه سازی توزیع ترافیک، مدیریت منابع و تخصیص بهینه آن‌ها، داده‌های کلیه گره‌های[1] اصلی شبکه مادر مخابراتی شرکت ارتباطات زیرساخت(به عنوان زیرساخت کلیه شبکه‌ها و خدمات کشور از جمله شبکه ملی اطلاعات) را پایش کرده و اطلاعات آن‌ها را جهت ارائه گزارش‌های تحلیلی سطح بالا از وضعیت شبکه، مورد داده کاوی قرار دهد. بدین منظور برای پایش وضع موجود و پیش بینی الزامات توسعه آتی شبکه زیرساخت، نیاز به سامانه‌ای است که توانایی پایش مداوم کلیه سیستم‌ها را به صورت خودکار دارا باشد. لذا می‌توان یک «سامانه جمع‌آوری، پردازش، تحلیل، ذخیره و نمایش اطلاعات و آمار ترافیکی شبکه» ایجاد نمود تا به پایگاه داده های سیستم های مدیریت موجود در شبکه و سیستم های پایش متصل شده و با بهره گیری از الگوریتم‌های هوش مصنوعی به پایش و ارائه تحلیل و پیش بینی جامعی از کل شبکه بپردازد.

این گزارش با هدف ارائه راه‌کارها و مفاهیم گردآوری فیلدهای داده‌ای از لاگ ترافیک شبکه، محاسبه پارامترهای ترافیکی و تحلیل آماری و توصیفی هوشمند آن‌ها ارائه شده است که هسته پژوهشی این گزارش و برخی مثال‌ها و نتایج ارائه شده در آن مبتنی بر پژوهش‌های پایه‌ای و توسعه عملی انجام شده در دو پروژه « تحلیل نیازمندی‌ها، طراحی و پیاده‌سازی سامانه بومی رتبه‌بندی و پایش وب‌سایت‌ها » و « طراحی و ایجاد بستر تحلیل ترافیک و ذائقه سنجی کاربران فضای مجازی» که در سال‌های اخیر در گروه مدیریت یکپارچه شبکه انجام شده‌اند، می‌باشد. بدین منظور در این گزارش ابتدا به بیان مفاهیم داده‌ها در لایه‌های شبکه و پروتکل‌های ارتباطی مختلف و انواع فیلدهای داده‌ای قابل استخراج از طریق آن‌ها پرداخته شده و سپس راه‌کارهایی جهت استخراج و گردآوری داده‌ها ارائه شده است. در بخش آخر گزارش نیز به معرفی مفاهیم و ارائه راهکارهای تحلیل داده‌ها به منظور مدیریت و پایش شبکه پرداخته شده است.

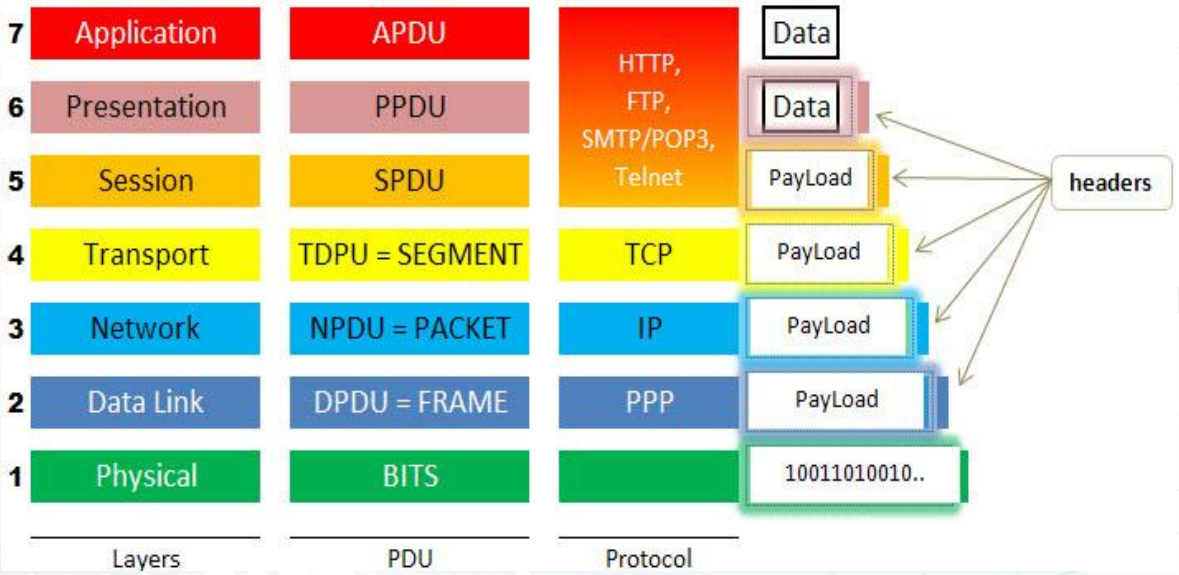
[1] Nodes

۲ داده‌های مورد نیاز برای پایش ترافیک شبکه

داده‌های مورد نیاز برای تحلیل از تجهیزات شبکه در لایه‌های مختلف شبکه و در نقاط مختلف آن استخراج می‌شوند. فیلدهای داده‌ای مورد نیاز برای پایش و تحلیل شبکه با نیازها و سیاست‌های تحلیل ارتباط مستقیم دارد. فیلدهای داده‌ای از بسته‌ها و پیام‌ها در لایه‌های مختلف شبکه استخراج می‌شوند که آن‌ها نیز در قالب پروتکل‌های مشخص انتقال می‌یابند. پروتکل‌های انتقال فیلدهای داده‌ای مشخصی دارند که در هنگام ارتباط و انتقال پیام مقداردهی می‌شوند؛ از این فیلدها برای پایش و تحلیل ترافیک شبکه‌ها می‌توان استفاده نمود. فیلدهای داده‌ای لایه کاربرد بیشترین و کاربردی‌ترین اطلاعات تحلیل را در اختیار کاربران قرار می‌دهد. در این بخش ضمن یادآوری مفاهیم لایه‌های شبکه، فیلدهای داده‌ای پروتکل‌های مطرح در لایه کاربرد معرفی شده‌اند و سپس فیلدهای داده‌ای محاسباتی که از روی آن‌ها قابل ارائه است معرفی شده است.

۲-۱ لایه‌های شبکه و انتقال داده‌ها

منابع داده‌ای لاگ شبکه، بر اساس جریان ترافیکی عبوری از لایه‌های مختلف شبکه از مبدأ (دستگاه کاربر) تا مقصد (سرور دامنه) تشکیل می‌گردد. در شبکه‌های کامپیوتری به‌طورکلی یک مدل عمومی برای عبور جریان داده‌ها از مبدأ تا مقصد وجود دارد. شکل ۱ لایه‌های مختلف شبکه بر اساس مدل OSI و واحد انتقال داده متناظر با هرکدام را نشان می‌دهد[۱].



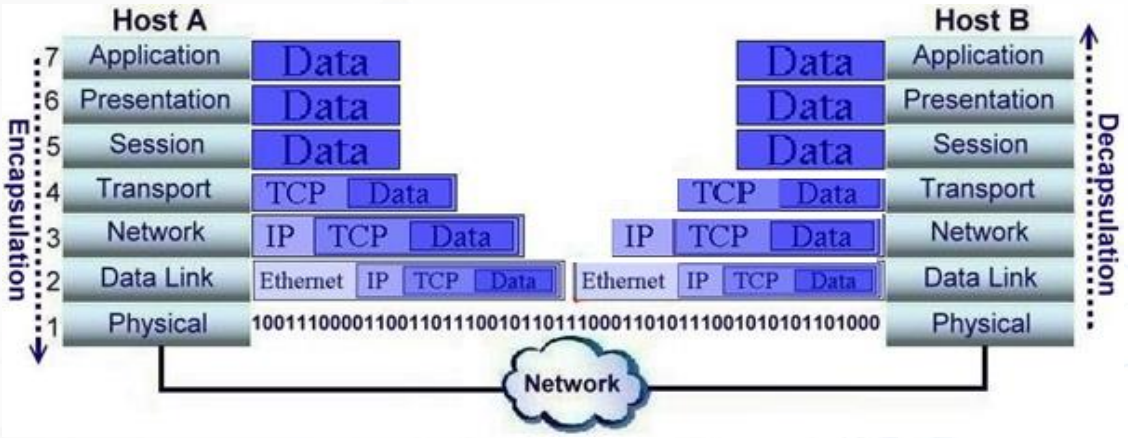
شکل ۱: لایه‌های شبکه در مدل مرجع OSI و واحدهای انتقال اطلاعات در هر لایه

بر اساس این مدل هر بسته ارسالی در شبکه از ترکیب payload+header تشکیل می‌شود و یک واحد انتقال بسته در شبکه را می‌سازد. از مهم‌ترین فیلدهای موجود در این سرآیندها که در پایش ترافیک شبکه موثر هستند، می‌توان به موارد ذیل اشاره نمود:

- IP مبدأ و IP مقصد (سرآیند IP)
- PORT مبدأ و PORT مقصد، زمان و تاریخ درخواست (سرآیند TCP)

- نام دامنه، آدرس URL، شناسه عامل کاربری (User-agent)، نوع محتوا (Content-type) – (سرآیند پروتکل HTTP)

واحد انتقال داده در هر لایه از شبکه دارای یک نام است که در شکل۱ به آن اشاره شده است. واحد انتقال اطلاعات در هر لایه در فرآیندی موسوم به کپسوله کردن و از کپسوله خارج کردن بسته‌ها مطابق شکل ۲ صورت می‌پذیرد. در سمت مبدأ هر لایه، اطلاعات سرآیند خود را به بسته اضافه می‌کند، در سمت مقصد سرآیند اطلاعاتی متناظر با هر لایه جدا شده تا در بالاترین لایه(لایه کاربرد) به برنامه یا خدمت تحویل داده شود.



شکل ۲: فرآیند کپسوله سازی و از کپسوله خارج کردن در لایه‌های مختلف شبکه

۲-۲ جمع‌آوری داده‌های ترافیک شبکه

داده‌های ترافیک شبکه از لایه IP یا لایه سوم شبکه تا لایه کاربرد جهت تحلیل ترافیک و پایش شبکه و رفتارهای کاربران می‌تواند مورد استفاده قرار بگیرند. بنابراین جمع‌آوری داده‌ها نیز به روش‌های مختلفی در لایه‌های مختلف قابل انجام است. جمع‌آوری داده‌ها اغلب با استخراج لاگ فایل‌ها صورت می‌پذیرد. اطلاعاتی که لاگ در اختیار می‌گذارد، یک منبع بسیار مهم برای کسب دانش تحلیل است. به عنوان مثال با استفاده از لاگ یک سرور وب، به راحتی می‌توان تشخیص داد که چه تعداد کاربر و از چه موقعیت‌های جغرافیایی به فایل خاصی دسترسی داشته‌اند و یا با تجمیع لاگ‌های مربوط به ترافیک تبادلی تجهیزات شبکه می‌توان تشخیص داد که چطور می‌توان توزیع بار ترافیکی یا مدیریت منابع را بهینه ساخت و یا حتی پیش‌بینی کرد که در آینده ممکن است چه خطراتی برای شبکه وجود داشته‌باشد. لازم به ذکر است که بهره‌گیری از انواع داده‌ها یا انواع روش‌ها و یا نقاط جمع‌آوری داده در معماری شبکه، همگی وابسته به نیاز و سیاست‌های تحلیل و پایش ترافیک شبکه است و در برخی موارد ممکن است از چندین روش به صورت ترکیبی برای جمع‌آوری داده‌های یک شبکه استفاده نمود.

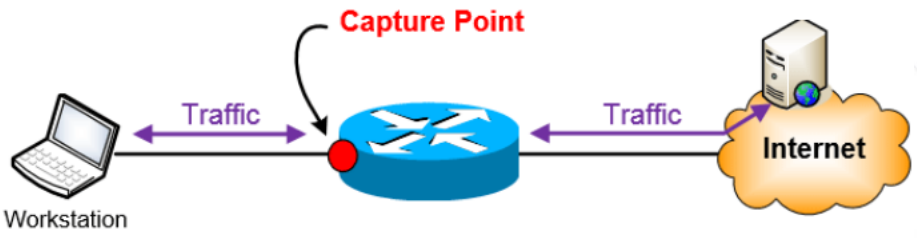
پایش و تحلیل ترافیک شبکه را می‌توان از دو منظر بررسی نمود که یکی پایش ترافیک شبکه به منظور تحلیل و بررسی رفتار شبکه و کمک به مدیریت هوشمند شبکه بوده و دیگری پایش ترافیک وب و کاربردها در شبکه به منظور تحلیل رفتار کاربران شبکه و کمک به شناخت بهتر گرایشات کاربران و در نهایت سیاست گذاری دقیق‌تر در این راستا است. با در نظر گرفتن این دو دیدگاه و اهداف و نیازهای مدیریتی در هریک و همچنین روش‌ها و نقاط مختلف گردآوری داده از ترافیک شبکه‌ها، فیلدهای داده‌ای مورد نیاز انتخاب می‌شوند که در ادامه به شرح آن‌ها با در نظر

گرفتن ماهیت شبکه‌های اپراتورهای فراهم کننده خدمات اینترنت و اینترانت به عنوان زیرساخت شبکه ملی اطلاعات، پرداخته شده است.

۲-۲-۱جمع‌آوری لاگ از نودهای سوئیچ و روتر شبکه

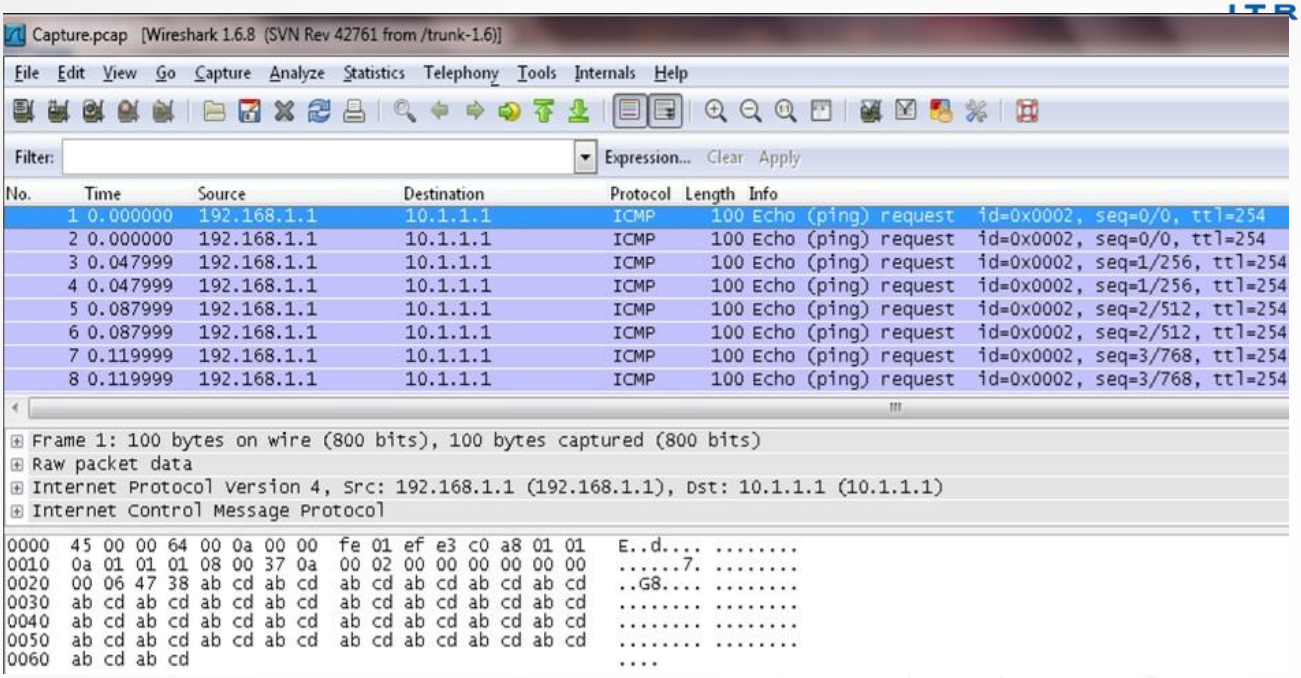
تمام گره‌های شبکه جهت اتصال به شبکه از واسط کارت شبکه استفاده می‌کنند. سوئیچ‌ها، روترها و دروازه‌ها که با کارکرد اصلی توزیع و مسیریابی ترافیک از مبدأ به مقصد مورد استفاده قرار می‌گیرند، مبادی ورودی و خروجی شبکه با توجه به طرح معماری شبکه هستند. بنابراین استخراج لاگ این گره‌های مهم شبکه که از لایه ۳ تا ۷ شبکه می‌توانند داده‌ها را انتقال دهند(اغلب در لایه ۳ فعالیت دارند)، جهت تحلیل و پایش ترافیک شبکه، داده‌های مفید و موثری را در اختیار می‌گذارند. نحوه استخراج لاگ از این تجهیزات می‌تواند برای برخی از آن‌ها که قابلیت لاگ گیری دارند، استفاده از لاگ آن‌ها باشد و برای برخی دیگر با قرار دادن یک سرور پشتیبان و ارسال یک نسخه از کلیه ترافیک عبوری به آن صورت پذیرد. البته پروتکل‌ها و ابزارهای مدیریت شبکه نیز همانطور که بخش مربوطه ذکر شده است، می‌تواند در دریافت ترافیک این تجهیزات کمک کنند.

از ابزار Wireshark برای پایش تمام ترافیک عبوری بر روی کارت شبکه استفاده می‌شود. هدف معرفی این ابزار در این گزارش جمع‌آوری و بررسی اطلاعات مربوط به ترافیک وب‌سایت‌ها و URL های مورد بازدید توسط کاربران شبکه و شناسایی نحوه رفتار و تبادل همه بسته‌های تبادلی است. در شکل ۳ نقطه لاگ گیری نشان داده شده است. نقطه لاگ گیری از طرح معماری شبکه کاملا وابسته به اهداف و سیاست‌های لاگ‌گیری است.



شکل ۳: نقطه لاگ گیری از گره‌های سوئیچ و روتر در شبکه

همان‌طور که در شکل ۴ مشاهده می‌شود، بسته‌هایی که از روتر گذر کرده‌اند، با حداقل سه فیلد مهم **زمان**، **آدرس IP مبدأ** و **آدرس IP مقصد** لاگ می شوند. البته اگر قابلیت لاگ گیری بالاتر از لایه ۳ شبکه وجود داشته باشد، فیلدهای داده‌ای بیشتری برای لاگ گرفتن وجود خواهد داشت. با این حال با این ۳ فیلد نیز می‌توان وضعیت درخواست‌ها، ترافیک شبکه، آمار بازدیدها و این قبیل معیارها را محاسبه نمود و در واحد تحلیل مورد استفاده قرار داد.



شکل ۴: نمونه لاگ سوئیچ و روترهای لایه ۳

۲-۲-۲جمع آوری لاگ از سیستم‌های مدیریت شبکه

پروتکل‌های مدیریت شبکه به همراه ابزارها و سیستم‌های مدیریت شبکه، داده‌ها و اطلاعات زیادی از وضعیت شبکه را در اختیار تحلیلگران داده‌های ترافیک شبکه قرار می‌دهند. در ادامه به شرح مختصری از برخی از پرکاربردترین این پروتکل‌ها پرداخته شده است.

۱-۲-۲-۲ پروتکل Syslog

تجهیزات شبکه از پروتکل Syslog برای ارسال پیغام رخداد های خود به یک سرور گیرنده پیغام که اصطلاحا Syslog Serverنامیده می شود استفاده می کنند. این پروتکل کلاینت-سروری توسط بسیاری از انواع تجهیزات شبکه مانند سوئیچ ها، روترها و فایروال ها پشتیبانی می شود. تجهیزات شبکه می توانند به گونه ای تنظیم شوند تا رخداد های مختلفی را با ارسال پیغام Syslog اطلاع رسانی کنند، برای مثال یک روتر ممکن است هر بار که یک کاربر به کنسول آن وارد شد پیغام Syslog ارسال کند، و یک وب سرور زمانی که رخداد access-denied روی داد، پیغام Syslog ارسال کند. بیشتر تجهیزات شبکه مانند سوئیچ ها و روتر ها می توانند پیغام Syslog ارسال کنند، علاوه بر اینها، همه سیستم عامل های مبتنی بر Unix و همه انواع سیستم عامل های Linux از Syslog به عنوان روش اصلی مدیریت لاگ استفاده می کنند. بسیاری از برنامه های کاربردی و خدمت ها مانند Oracle و Apacheنیز از این پروتکل پشتیبانی می کنند. استفاده از Syslog روشی مناسب برای جمع آوری و متمرکز سازی لاگ های رخدادها اعم از خطاها، هشدارها و خرابی‌ها از همه تجهیزات شبکه است[۲].


پنل مشاهده Syslog

پیغام	زمان ارسال	Severity	Facility	فرستنده
User Level Messages Alerts	1392/02/09 14:22:0	Warning	System	192.168.3.22
User Level Messages Alerts	1392/02/09 14:21:58	Error	System	192.168.3.22
User Level Messages Alerts	1392/02/09 14:21:56	Critical	System	192.168.3.22
User Level Messages	1392/02/09 14:21:54	Alert	System	192.168.3.22

شکل ۵ : نمونه لاگ پروتکل Syslog با استفاده از یک ابزار مانیتورینگ

همان‌طور که در شکل ۵ دیده می‌شود، فیلدهای داده‌ای که از این پروتکل جمع‌آوری می‌شود شامل مواد زیر است:

- Hostname: نام و یا آدرس IP سیستم ارسال کننده پیغام Syslog
- Timestamp: زمان ایجاد پیغام .Syslog
- Message: متن پیغام Syslog که معمولا توضیحاتی در خصوص رخداد مربوطه ارائه می کند.

۲-۲-۲-۲ پروتکل SNMP

یکی از پروتکل های مدیریت شبکه SNMP[۱] به معنی پروتکل مدیریت ساده شبکه می باشد. تجهیزات شبکه همچون هاب ها سوییچ ها و روتر ها دارای SNMP Agent هستند که این عامل ها یا Agent ها اطلاعات تجهیز را با استفاده از درگاه ۱۶۱ UDP جمع آوری می کنند سپس این اطلاعات را به SNMP Managers یا همان نرم افزار های مدیریت SNMP ارسال می کنند. این نرم افزار ها برای جمع آوری اطلاعات و بررسی و تحلیل وقایع شبکه طراحی شده اند.

به طور کلی می‌توان گفت SNMP پروتکل مدیریت پهنای باند است که مواردی مانند لودCPU ، مصرف رم و وضعیت اینترفیس را بررسی می‌کند. بنابراین مدیران شبکه جهت نظارت بر پهنای باند از داده‌های لاگ این پروتکل استفاده می‌کنند؛ هرچند امروزه نسخه پیشرفته این پروتکل تحت عنوان پروتکل NetFlow مورد استفاده مدیران شبکه است. NetFlow برای شبکه‌های پیچیده با ترافیک بالا کاربرد دارد تا آنومالی شبکه تشخیص داده شود[۳].

۲-۲-۲-۳ پروتکل NetFlow [۴]

تحلیلگر NetFlow که نرم افزار مدیریت شبکه مبتنی بر Flow است، با NetFlow و sFlow و jFlow و IPFIX و دیگر فرمت‌های Flow یکپارچه است تا Visibility بلادرنگ از ترافیک شبکه و کارایی پهنای باند ارایه دهد . NetFlow Analyzerبه تشخیص و عیب یابی کندی و آنومالی شبکه به همراه جزییات دقیق کمک می‌کند و نیازهای پهنای باند شما در آینده را مشخص می‌کند. از جمله قابلیت‌های NetFlow عبارتند از:

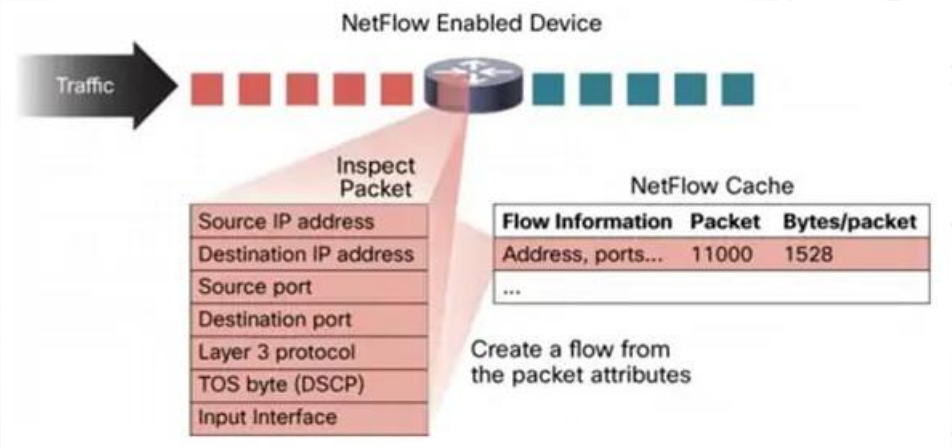
- آنالیز برنامه‌های جدید و تاثیر آنها روی شبکه

[۱] Simple Network Management Protocol

- کاهش پیک ترافیکWAN
- عیب یابی و فهمیدن Pain Point های شبکه
- تشخیص ترافیک غیرمجازWAN
- امنیت و تشخیص آنومالی شبکه
- تایید اعتبار پارامترهایQoS – Quality of Service
- ارایه گزارشات
- کاهش زمان عیب یابی
- رفع مشکلات گلوگاه

هر بسته‌ای که در سوئیچ و روتر، فوروارد می‌شود توسط NetFlow از نظر مجموعه‌ای از ویژگی‌های بسته IP بررسی می‌شود. این ویژگی‌ها عبارتند از: هویت بسته IP یا همان اثر انگشت بسته و تشخیص اینکه بسته منحصر به فرد است یا شبیه بسته های دیگر.

تمام بسته‌هایی که آدرس IP مبدا / مقصد، درگاه مبدا / مقصد، protocol interfaceو CoS یکسانی دارند در یک flow گروه‌بندی و سپس بررسی می‌شوند. حجم زیادی از اطلاعات شبکه در دیتابیس اطلاعات NetFlow مختصر می‌شوند. به این دیتابیس NetFlow Cache می‌گویند. بنابراین روش اثرانگشت و تشخیص flow ارتقاپذیر است.



شکل ۶: فیلدهای داده‌ای پروتکل NetFlow

IP Flow مجموعه‌ای با ۵ تا ۷ ویژگی بسته IP است. فیلدهای داده‌ای مربوط به بسته IP که NetFlow استفاده و لاگ می‌کند، در ذیل آورده شده است. این داده‌های Flow برای تحلیل رفتار شبکه بسیار مهم است:

- آدرس مبدا که مشخص می‌کند چه کسی ترافیک را ایجاد می‌کند.
- آدرس مقصد که مشخص می‌کند چه کسی ترافیک را دریافت می‌کند.
- درگاه‌ها که مشخص می‌کند چه اپلیکیشن‌هایی از ترافیک شبکه استفاده می‌کنند.
- CoS که اولویت ترافیک را بررسی می‌کند.
- اینترفیسِ دستگاه مشخص می‌کند دستگاه شبکه چگونه از ترافیک استفاده کند.
- بایت‌ها و بسته های بررسی شده، مقدار ترافیک را مشخص می‌کنند.

- Timestampکه مشخص کننده عمر Flow است timestamp .برای محاسبه بایت‌ها و بسته‌ها در هر ثانیه به کار می‌رود.
- Next hop IP addressها که شامل BGP routing Autonomous Systems (AS) است.
- Subnet maskدر آدرس های مبدا و مقصد برای محاسبه پیشوندها
- TCP flagبرای بررسیTCP handshake

۲-۲-۳جمع‌آوری داده‌های ترافیک وب

تحلیل و بررسی ترافیک وب ابزاری قدرتمند جهت استخراج الگوهای رفتاری کاربران و پیش‌بینی گرایش و روالِ علاقه‌مندیِ آن‌ها در آینده بوده است. نیاز به استخراج فیلدهای داده‌ای و محاسبه پارامترهای ترافیکی بر اساس اطلاعات منابع داده‌ای (لاگ) می‌باشد. این پارامترها، به‌صورت شناخته‌شده‌ای در حوزه علمی تجزیه‌وتحلیل وب[۱] بوده و دارای تعاریف استانداردی می‌باشند. مطرح‌ترین پروتکل‌های لایه کاربرد جهت تبادل پیام‌های تحت وب، پروتکل‌های HTTP و HTTPS هستند. **در ادامه فیلدهای داده‌ای قابل استخراج از سرآیند پیام‌های این پروتکل‌ها آورده شده است. همچنین مثال‌ها و نمونه‌هایی از روش‌ها و نقاط مختلف لاگ‌گیری از شبکه‌های فراهم کنندگان خدمات اینترنت و اینترانت به عنوان زیرساخت شبکه ملی اطلاعات، شرح داده شده است.**

۲-۲-۳-۱ پروتکل‌های اصلی وب(HTTP و HTTPS)

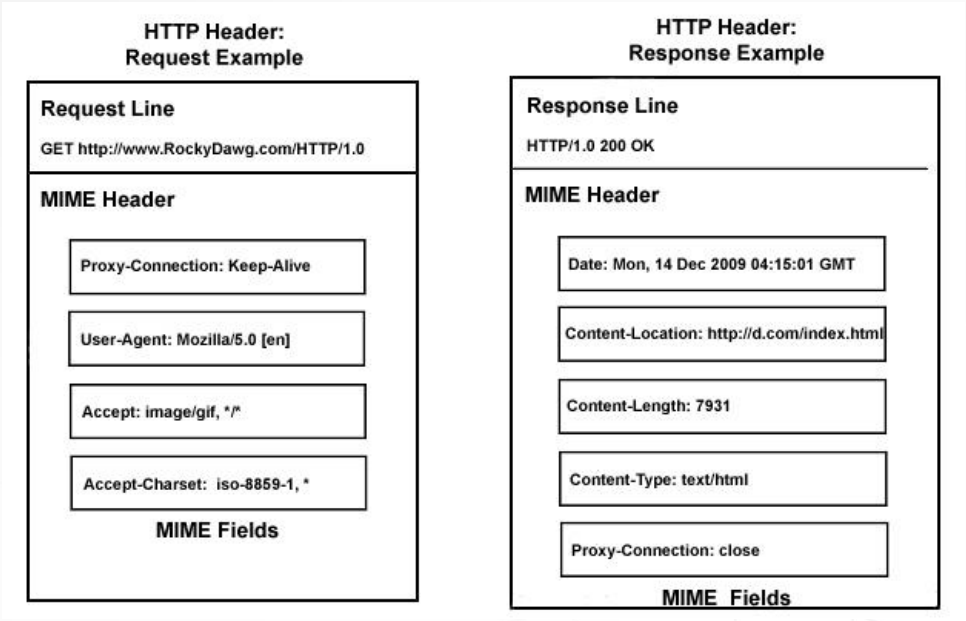
بازدید از وب توسط کاربران اغلب توسط پروتکل‌های HTTP و HTTPS انجام می‌شود و فیلدهای مهم در لاگ ترافیک مراکز داده آدرس مبدأ و آدرس وب یا URL مقصد است که در کنار زمان درخواست، مدت زمان نشست، عامل کاربری[۲] و سایر فیلدهای داده‌ای در سرآیند این پروتکل‌ها، اطلاعات تحلیلی زیادی را در اختیار قرار می‌دهند که در ادامه به معرفی آن‌ها پرداخته شده است. با این حال در برخی روش‌های گردآوری داده‌های ترافیک وب از پروتکل‌های دیگری مانند DNS استفاده می‌شود که در بخش‌های بعدی شرح داده شده‌اند.

در شکل ۷ نمونه‌ای از سرآیند پیام‌های انتقالی توسط پروتکل HTTP از مبدأ به مقصد، نشان داده شده است. سرآیند این پیام مربوط به لایه کاربرد است و فیلدهای داده‌ای در سرآیند این پیام‌ها مقداردهی شده‌اند[۵].

---

۱ Web Analytics

۲ User Agent(OS/Browser/Devoce/…)



شکل۷: سرآیند بسته‌های پروتکل HTTP

جدول ۱ ، فیلدهای ترافیک قابل جمع‌آوری از درخواست‌های HTTP که در لاگ ترافیک شبکه ثبت می‌شوند را نشان داده است[۵].

جدول ۱: فیلدهای ترافیک درخواست‌های HTTP در لاگ شبکه

ردیف	نام فیلد	توضیح
۱.	total_len	حداکثر طول فریم(واحد داده در لایه پیوند داده/اترنت) برحسب بایت به همراه سرآیند و پس آیند (MTU+header+trailer)[۱]
۲.	start_time	زمان شروع ارتباط(درخواست) بر روی تجهیز
۳.	cdr_id	شناسه رکورد جزئیات تماس(درواقع CDR Session id). یک ویژگی تکمیلی برای رکورد جزئیات تماس یا CDR[۲] است که اطلاعات اضافه‌ای را درباره رکوردهای جزئیات تماس در دروازه های انتقال صوت و تلفن تحت کنترل سیسکو[۳] Cisco Unified CME یا Cisco Unified SRST ثبت می‌کند. این فیلد حاوی یک شناسه همبسته سنجی یکتا است که مشخصه تماس در بین همه legs های موجود در تماس را مشخص می‌کند[۴].
۴.	trans_time_ms	زمان سپری شده بین شروع و پایان ارتباط. زمان بین شروع ار سال اولین بیت و اتمام آخرین بیت اطلاعات.
۵.	user_ip	آدرس آی‌پی مبدأ ارسال‌کننده درخواست را نشان می‌دهد.

[۱] Catalyst ۴۵۰۰/۴۰۰۰ with Supervisor V

[۲] Call Details Record

[۳] Cisco IOS voice gateways

[۴] اطلاعات رکورد جزئیات تماس می توانند درقالب خروجی پیام های لاگ سیستم(Syslog) یا ارسال به یک RADIUS VSA مورد استفاده قرار گیرند.

ردیف	نام فیلد	توضیح
۶.	Apn	Access Point Name یا نام نقطه دسترسی مشخص‌کننده دروازه بین شبکه GPRS، GSM، نسل ۳ یا ۴ و یا شبکه رایانه‌ای(شبکه حامل) با اینترنت است. APN شبکه داده بسته‌ها(PDN) و نوع خدمت ارائه شده توسط شبکه (نظیر MMS، WAP) را معین می‌کند. هنگامی یک دستگاه موبایل قصد برقراری ارتباط داده را دارد باید APN شبکه حامل خود را تنظیم نماید. مثلاً در تنظیمات گوشی برای شبکه همراه اول MCINET تنظیم می‌شود.
۷.	dest_ip	آدرس IP مقصد را مشخص می‌کند.
۸.	user_port	شماره درگاه مبدأ ارتباط را مشخص می‌کند.
۹.	dest_port	شماره درگاه مقصد ارتباط را مشخص می‌کند.
۱۰.	up_packets	تعداد بسته ارسالی را تعیین می‌نماید.
۱۱.	down_packets	تعداد بسته دریافتی را تعیین می‌نماید.
۱۲.	up_bytes	حجم داده ارسالی(به بایت) را مشخص می‌کند.
۱۳.	down_bytes	حجم داده دریافتی(به بایت) را مشخص می‌کند.
۱۴.	protocol_type	نوع پروتکل مورداستفاده را تشریح می‌کند. برای این لاگ درصورتی‌که مقدار معادل ۷ باشد بیانگر پروتکل HTTPS (درگاه ۴۴۳) و اگر مقدار ۴۰ باشد بیانگر استفاده ا از درگاه غیر HTTP است(درگاه غیر از ۸۰)[۱]
۱۵.	syn_synack_delay	تأخیر بین بسته syn و synack در زمان برقراری ارتباط(مربوط به‌طرف فرستنده در یک ارتباط) را مشخص می‌کند.
۱۶.	synack_ack_delay	تأخیر بین بسته synack و ack در زمان برقراری ارتباط(مربوط به‌طرف گیرنده در یک ارتباط) را مشخص می‌کند.
۱۷.	ack_first_data_delay	تأخیر دریافت ACK مربوط به اولین داده ارسالی را مشخص می‌کند.
۱۸.	first_second_data_delay	تأخیر زمانی بین اولین و دومین داده دریافتی مربوط به یک ات صال را مشخص می‌کند.
۱۹.	req_count	بیانگر تعداد درخواست‌های ردوبدل شده در یک ارتباط است.
۲۰.	content_length	طول محتوای بسته در ارتباط را نشان می‌دهد.
۲۱.	resp_delay	N/A
۲۲.	dl_delay	N/A
۲۳.	last_ack_delay	تأخیر زمانی در دریافت ACK مربوط به آخرین داده ارسالی در یک اتصال را مشخص می‌کند.
۲۴.	Version	N/A
۲۵.	first_page_flag	پرچم صفحه اول. به ازای بازدید صفحه خانگی یک منبع مقدار ۱ می‌گیرد.
۲۶.	user_agent	نرم‌افزار یا عامل نرم‌افزاری که فرآیند بازیابی محتوای وب را انجام دهد.
۲۷.	Uri	رشته‌ای از کاراکترها که به‌عنوان نشانگر منابع بکار برده می‌شود

[۱] این تحلیل از روی بررسی لاگ به صورت شهودی آورده شده است و ممکن است بسته به مورد استفاده تفسیر متفاوت داشته باشد.

ردیف	نام فیلد	توضیح
۲۸.	Host	نام دامنه که شامل خدمت‌دهنده دامنه همراه با شماره درگاه می‌باشد.
۲۹.	content_type	نوع محتوا، فرمت‌های قابل‌استفاده در دریافت و ارسال محتوا را مشخص می‌کند.

پروتکل HTTPS همان پروتکل HTTP رمزنگاری شده با استفاده از پروتکل TLS یا SSL است که امنیت بیشتری در ارتباطات ایجاد نموده و در عین حال محدودیت در دستیابی به داده‌های ارتباط در سرآیند آن در لاگ شبکه وجود دارد. جدول ۲، فیلدهای ترافیک قابل جمع‌آوری از درخواست‌های HTTPS که در لاگ شبکه ثبت می‌شود، را ارائه کرده است.

جدول ۲: فیلدهای ترافیک درخواست‌های HTTPS در لاگ شبکه

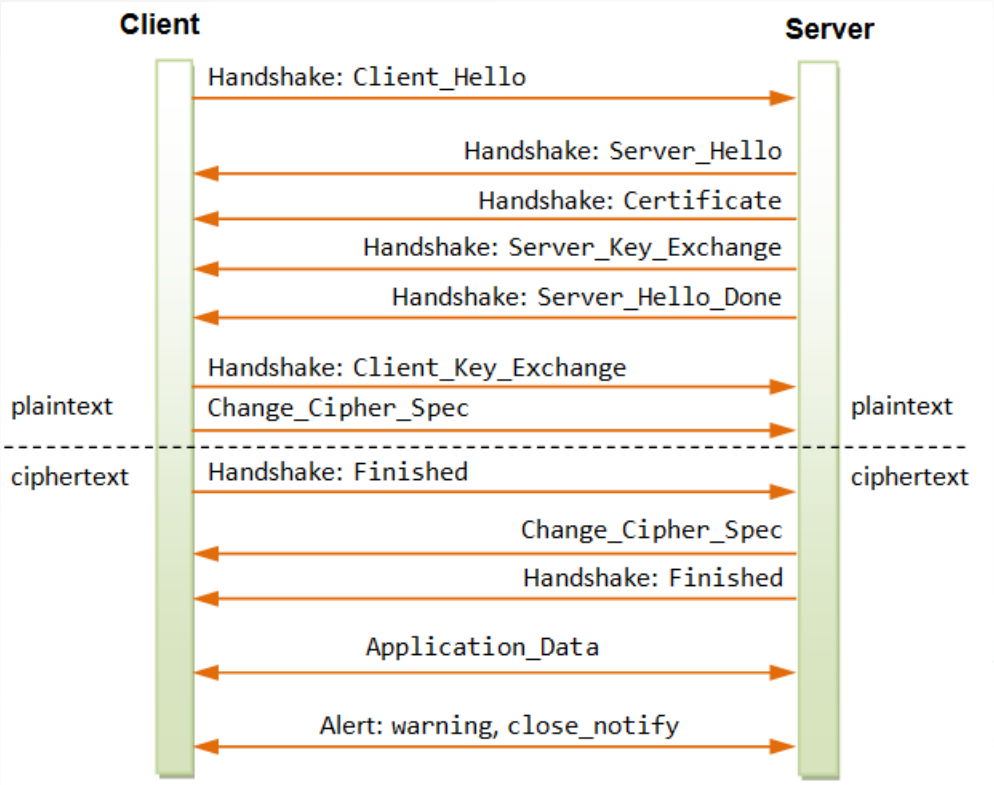
ردیف	نام فیلد	توضیح
۱.	total_len	حداکثر طول فریم(واحد داده در لایه پیوند داده/اترنت) برحسب بایت به همراه سرآیند و پس آیند (MTU+header+trailer)
۲.	Table_id	N/A
۳.	Server_detail	اطلاعات مرتبط با سرور ارسال‌کننده/دریافت‌کننده لاگ
۴.	start_time	زمان شروع ارتباط(درخواست) بر روی تجهیز
۵.	Trans_time	زمان سپری شده بین شروع و پایان ارتباط. زمان بین شروع ار سال اولین بیت و اتمام آخرین بیت اطلاعات.
۶.	Sgsn_u_ip	آدرس IP، SGSN (گره پایه خدمات) در شبکه GSM
۷.	Ggsn_u_ip	آدرس IP، GGSN (گره پایه Gateway) در شبکه GSM
۸.	dest_ip	آدرس IP مقصد را مشخص می‌کند.
۹.	user_port	شماره درگاه مبدأ ارتباط را مشخص می‌کند.
۱۰.	dest_port	شماره درگاه مقصد ارتباط را مشخص می‌کند.
۱۱.	Tcp_fin_pkts	نشان‌دهنده‌های تعداد بسته‌های FIN دریافت شده
۱۲.	up_tcp_dropped_pkts	تعداد بسته‌های TCP ارسالی ازدست‌رفته
۱۳.	down_tcp_dropped_pkts	تعداد بسته‌های TCP دریافتی ازدست‌رفته
۱۴.	up_packets	تعداد بسته ارسالی را تعیین می‌نماید.
۱۵.	down_packets	تعداد بسته دریافتی را تعیین می‌نماید.
۱۶.	up_bytes	حجم داده ارسالی (به بایت) را مشخص می‌کند.
۱۷.	down_bytes	حجم داده دریافتی(به بایت) را مشخص می‌کند.
۱۸.	tcp_window_size	اندازه پنجره TCP
۱۹.	proto_type	نوع پروتکل (به دلیل حجم لاگ ها قابل تشخیص نیست)
۲۰.	segment_flag	پرچم مشخص‌کننده استفاده از قطعه‌بندی در لایه IP
۲۱.	tcp_syn_pkts	تعداد بسته‌های SYN
۲۲.	tcp_conn_succ	Flag مشخص‌کننده موفقیت‌آمیز بودن connection
۲۳.	tcp_syn_ack_pkts	تعداد بسته‌های SYN ACK

ردیف	نام فیلد	توضیح
۲۴.	tcp_ack_pkts	تعداد بسته‌های ACK
۲۵.	up_ip_frag_pkts	تعداد بسته‌های IP قطعه شده ارسالی
۲۶.	down_ip_frag_pkts	تعداد بسته‌های IP قطعه شده دریافتی
۲۷.	up_tcp_out_order_pkts	تعداد بسته‌های ارسالی خارج از نوبت را نشان دهد.
۲۸.	down_tcp_out_order_pkts	تعداد بسته‌های دریافتی خارج از نوبت را نشان دهد.
۲۹.	up_tcp_retrans_pkts	تعداد بسته‌های ارسالی تکراری
۳۰.	down_tcp_retrans_pkts	تعداد بسته‌های دریافتی تکراری
۳۱.	tcp_reset_pkts	تعداد بسته‌های RST
۳۲.	syn_synack_delay	تأخیر بین بســـته syn و synack در زمان برقراری ارتباط(مربوط به‌طرف فرستنده در یک ارتباط) را مشخص می‌کند.
۳۳.	synack_ack_delay	تأخیر بین بســـته synack و ack در زمان برقراری ارتباط(مربوط به‌طرف گیرنده در یک ارتباط) را مشخص می‌کند.

همان‌طور که در جداول فوق مشاهده می‌شود، برخی فیلدهای داده‌ای موجود در پروتکل HTTP مانند نام دامنه، از پروتکل HTTPS به دلیل رمزنگاری به سادگی استخراج داده از پروتکل HTTP نیست و به این منظور بایستی عملیات لاگ گیری فراتر از استفاده از لاگ سطح بالای سرورها و فایروال ها بوده و از لاگ شبکه در حالت دستکانی[۱] استفاده نمود که در ادامه به شرح این راه‌کار پرداخته شده است.

بدین منظور نیاز است تا برای استخراج فیلد پایه‌ای و مهم نام دامنه از افزونه SNI در TLS استفاده کرد. بررسی ارتباطات HTTPS نشان می‌دهد علیرغم استفاده از رمزنگاری، برخی اطلاعات در فرآیند دستکانی SSL قابل مشاهده هستند. در SSL قبل از شروع ارسال اطلاعات یک دستکانی بین سرور امن و کاربر درخواست‌کننده شروع ارتباطات انجام می‌گردد. اطلاعاتی که قبل از شروع انتقال امن داده بین طرفین تبادل می‌شوند به‌صورت متن آشکار هستند و رمز نمی‌شوند. در شکل ۸ فرآیند برقراری ارتباط SSL قبل از ارسال اطلاعات لایه کاربردی به تصویر کشیده شده است[۶].

[۱] HandShaking



شکل ۸: فرآیند دستکانی و برقراری ارتباط در HTTPS

کاربر پیام Client Hello را به سمت سرور امن ارسال می‌نماید. این پیام حاوی اطلاعات رمزنگاری نظیر نسخه SSL/TLS و ترتیبی از الگوریتم‌های رمزنگاری مورد قبول کاربر است. سرور امن با ارسال پیام Server Hello پاسخ کاربر را می‌دهد. سرور همچنین گواهینامه دیجیتال خود را ارسال می‌کند، بعدازاین مراحل تبادل کلید رمز بین کلاینت و سرور انجام می‌شود. با پایان مراحل دست تکانی و ارسال کلیدهای نشست، ارسال اطلاعات شروع می‌گردد.

در سال ۲۰۰۷ ، افزونه‌ای بنام شناسه نام سرور یا SNI[۱] به پروتکل TLS افزوده شد. قابلیت SNI در HTTPS این امکان را فراهم می‌آورد تا کاربر نام میزبان امن موردنظر خود را در زمان شروع فرآیند دست تکانی TLS مشخص نماید. این قابلیت به سرور میزبان امکان ارائه چندین گواهینامه بر روی یک آدرس IP را در درگاه TCP را می‌دهد. درنتیجه می‌توان چندین وب‌سایت امن(HTTPS) بر روی یک آدرس IP یکسان میزبانی کرد بدون آنکه نیاز باشد همه وب‌سایت‌ها از یک گواهینامه یکسان استفاده کنند. در این حالت سرور به هر کاربر گواهینامه متناسب با نام دامنه متناظر را ارائه می‌دهد. بدون استفاده از SNI، هر وب‌سایت امن نیازمند داشتن یک آدرس IP مجزا است و یا همه وب‌سایت‌های میزبانی‌شده روی یک سرور باید از یک گواهینامه یکسان استفاده نمایند.

وجود این قابلیت این امکان را فراهم می‌کند که اطلاعات دامنه مورد درخواست کاربر را بتوان از ترافیک شبکه استخراج کرد؛ ازآنجاکه بسیاری از مرورگرها و وب سرورهای کنونی از قابلیت SNI پشتیبانی می‌کنند و تنها درصد کمی از مرورگرهای وب قدیمی با این قابلیت سازگار نیستند، می‌توان از این رویکرد برای استخراج فیلدهای داده‌ای مورد نیاز تحلیل وب با استفاده از ترافیک به‌دست‌آمده از شبکه استفاده کرد.

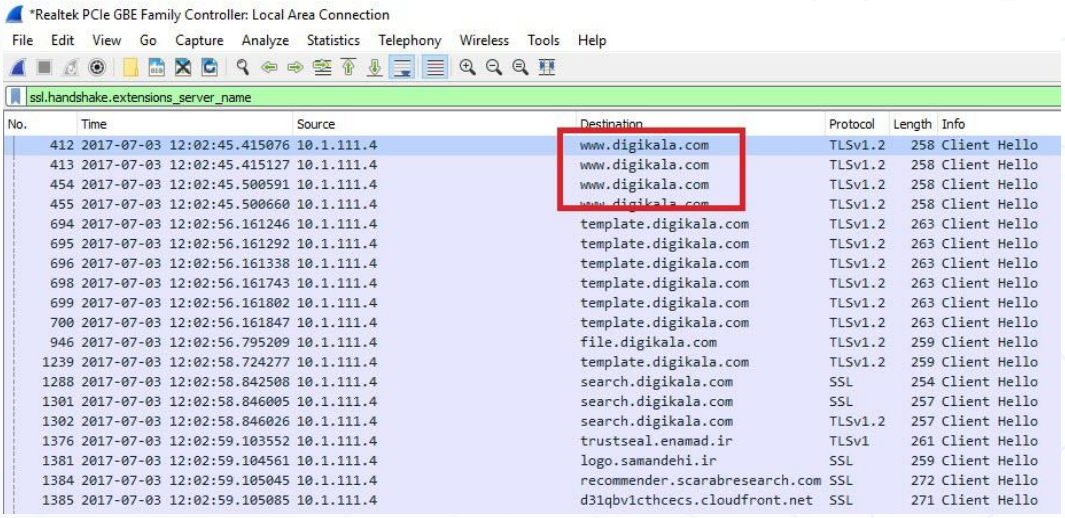
شکل ۹ نمایی از ترافیک لاگ شده برای یک وب‌سایت نمونه و مقدار فیلدهای SNI در آن را نشان می‌دهد.

[۱] Server Name Identification

```
Session ID: 6f140000393e0610a30fc6060bd2ee32ef79a0fac8715fc7...
Cipher Suites Length: 56
> Cipher Suites (28 suites)
Compression Methods Length: 1
> Compression Methods (1 method)
Extensions Length: 113
v Extension: server_name (len=17)
    Type: server_name (0)
    Length: 17
  v Server Name Indication extension
      Server Name list length: 15
      Server Name Type: host_name (0)
      Server Name length: 12
      Server Name: www.bing.com
v Extension: status_request (len=5)
    Type: status_request (5)
```

شکل ۹: مشاهده اطلاعات نام دامنه مورد درخواست کاربر از فیلد SNI

نمونه محتوای فیلد مربوط به نام سرور در افزونه SNI بر روی ترافیک ضبط‌شده در Wireshark در شکل ۱۰ نمایش داده شده است.



شکل ۱۰: استفاده از فیلتر Wireshark برای استخراج اطلاعات نام دامنه از ترافیک SSL/TLS

همچنین همان‌طور که در شکل ۱۰ دیده شد، از دیگر اطلاعاتی که در مرحله دست تکانی ssl در شبکه می‌توان استخراج نمود اطلاعات مربوط به نشست ssl کاربر است که در فیلد session ID قرار دارد. این شناسه تا زمانی که نشست کاربر با سرور برقرار است تغییر نمی‌کند. این شناسه بیانگر آن است که سرور خواستار استفاده از پارامترهای یکسان و مجموعه الگوریتم‌های رمزنگاری مورد توافق (مشخص‌شده در قسمت اول دست تکانی) است. هر زمان که سرور نخواهد این نشست را ادامه دهد یک بسته session id خالی به کاربر ارسال می‌کند. زمانی که سرور خواهان ادامه نشست با کاربر باشد این امر را از طریق ارسال پیام server hello با شناسه session id یکسان به کاربر اعلام می‌دارد.

همان‌طور که مشاهده می‌شود، فیلدهای داده‌ای زیادی در سرآیند پروتکل‌های وب قابل گردآوری هستند اما این سؤال مطرح است که آیا تمام آن‌ها برای تحلیل وب مورد نیاز است؟ در پاسخ به این سؤال باید گفت آنچه تعیین کننده است ابتدا اهداف و نیازهای تحلیلی سازمان و سپس محدودیت‌های آن در دسترسی یا ارائه داده‌ها است.

به عنوان مثال اگر هدف تحلیل آماری بازدیدهای کاربران(از نوع Hit) از دامنه‌های وب و از طریق پروتکل‌های مختلف باشد، فیلدهای داده‌ای زیر کفایت می‌کند:

- آدرس IP مبدأ
- آدرس Host Name یا URL یا URI
- زمان بازدید
- نوع پروتکل

در همین مثال ساده، نکاتی مطرح است که در انتخاب فیلدهای داده‌ای موثر واقع می‌شود. اولین نکته مربوط به **حفظ حریم خصوصی کاربران و محرمانگی** است؛ بدین معنا که هرگونه اطلاعاتی که به تنهایی و یا به کمک لاگ‌ها یا بانک‌های اطلاعاتی دیگر منجر به کشف هویت واقعی یک فرد شده و از طریق آن اطلاعات بتوان فعالیت‌های آن شخص را ردیابی نمود، نقض حریم خصوصی است و صرفا در شرایط خاص، مراجع ذی‌صلاح حق در اختیار داشتن این اطلاعات را دارند. بنابراین در این مثال آدرس IP کاربران که احتمال نقض حریم خصوصی را دارد نباید به طور مستقیم در پایگاه‌های داده‌ای ذخیره شده و مورد تحلیل قرار گیرد بلکه با روش‌هایی مانند نگاشت آدرس IP به یک ID منحصربفرد و یا هش نمودن آن، محرمانگی این فیلد داده‌ای در مراحل پردازش و تحلیل و نمایش حفظ شود.

نکته دیگر مربوط به پیش‌پردازش‌های لازم برای یک لاگ قبل از ورود آن به هر پایگاه داده‌ای جهت سایر پردازش‌ها و تحلیل است. در مثال فوق حذف خطاهای لاگ، آدرس‌های مبدأ و مقصد غیرمعتبر، فیلدهای خالی و شمارش بازدید به ازای هر آدرس URL یا URI از جمله پیش‌پردازش‌های مورد نیاز برای لاگ قبل از ورود به فرایند تحلیل است.

از این رو جهت انجام پیش‌پردازش‌های فوق شامل نگاشت IP به ID بهتر است به جای استفاده از لاگ خام، لاگی که پیش‌پردازشات لازم در آن انجام شده و اصطلاحا تمیز شده است، مورد استفاده قرار گیرد. برای مثال فوق یک فایل CSV با فیلدهای زیر را می‌توان متصور شد.

جدول۳: لاگ پردازش شده برای تحلیل آماری بازدید کاربران از دامنه‌های وب

Sheet name	**HTTP HIT**			
Columns names	Raw Lables			SUM of HIT
Raw Lables	domain	host	ID	Port
Sheet name	**SSL HIT**			
Columns names	Raw Lables			SUM of HIT
Raw Lables	domain	host	ID	Port

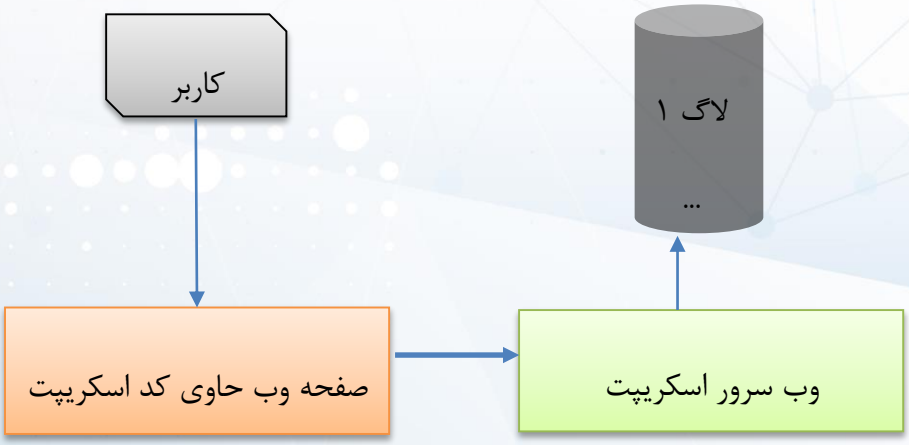
۲-۲-۳-۲ جمع‌آوری لاگ سرورهای میزبانی[۱] مراکز داده

یکی از نقاطی که لاگ ترافیک عبوری آن در پایش ترافیک شبکه‌ها، بالاخص شبکه اینترنت و اینترانت مفید و موثر است، لاگ سرورهای مراکز داده است. ترافیک کاربران شبکه های بزرگ جهت بازدید از وب به سوی اپراتورهای مراکز داده که میزبان سرورهای وب هستند، روانه می‌شوند؛ با بررسی ترافیک بازدید کاربران از دامنه‌های مختلف(با استفاده لاگ HTTP و HTTPS ایجاد شده) می‌توان اطلاعات ارزشمندی از گرایش کاربران در بازدید از آن‌ها و همچنین اطلاعات وضعیت شبکه و ترافیک آن بدست آورد.

۲-۲-۳-۳ جمع‌آوری لاگ ربات‌ها و اسکریپت‌ها

ربات‌ها، افزونه‌های مرورگرها و اسکریپت‌های وب یکی دیگر از روش‌های جمع‌آوری داده‌های ترافیکی بازدید از وب‌سایت‌ها است که در صورت حفظ جامعیت و تنوع می‌تواند در تحلیل ترافیک وب مورد استفاده قرار بگیرد. اسکریپت‌های وب توسط ابزارهای تحلیل وب مانند گوگل آنالیتیکز توسط صاحبان وبسایت‌ها روی صفحات وب ایشان قرار می‌گیرد و اطلاعات تحلیلی در اختیار آن‌ها می‌گذارد. همچنین ارگان‌های مرجعی در سراسر دنیا وجود دارد که اسکریپتی جهت ارزیابی فعالیت‌های وب به عنوان گواهینامه‌های مختلف به خواست صاحبان وب در وبسایت‌ها قرار می‌دهند که ترافیک آن وبسایت‌های دارای اسکریپت اطلاعات دقیق و سطح بالایی از بازدید کاربران از آن وب‌سایت‌ها را در اختیار تحلیل‌گران وب قرار می‌دهد. به عنوان مثال در ایران وزارت صمت اسکریپت نماد اعتماد الکرونیکی را ارائه می‌دهد و وزارت فرهنگ و ارشاد اسلامی اسکریپت نماد رسانه‌های دیجیتال را که یک ارزیابی محتوایی است، ارائه می‌دهد. در ادامه به شناخت بیشتری از لاگ‌های اسکریپتی پرداخته شده است.

فرآیند نحوه جمع‌آوری داده‌ها توسط اسکریپت در سطح انتزاعی مطابق شکل ۱۱ می‌باشد. اسکریپت مورد نظر توسط صاحبان وب‌سایت، در صفحات وب‌سایت جاسازی یا Embeded می‌شود، سپس کاربران بازدیدکننده وب‌سایت با مشاهده هریک از صفحات وب‌سایت، درخواستی برای بازدید از فایل اسکریپت نیز ارسال می‌نمایند، لذا این درخواست‌ها به وب سرور اسکریپت منتقل‌شده و بدین‌صورت اطلاعات ترافیکی بازدیدکنندگان وب‌سایت‌ها جمع‌آوری می‌گردد.



شکل ۱۱: دیاگرام چگونگی جمع‌آوری لاگ اسکریپت

[۱] Hosting

لاگ ترافیک دسترسی به وب سرور یا Web Server Access Log که اغلب توسط سرورهای وب تولید می‌شود، می‌تواند قالب‌های مختلفی داشته باشد. بر اساس وب سرورهای رایج امروزه سه نوع اصلی لاگ سرور وب عبارت‌اند از CLF[1]، Apache Custom Log Format و W3C Exended Log File. در جدول ۴ انواع فایل‌های رایج به همراه وب سرور و نمونه لاگ تولیدشده از آن‌ها نمایش داده می‌شوند.

جدول ۴: فرمت‌های رایج لاگ دسترسی به وب سرور

فرمت لاگ	نام سرور وب	نمونه رکورد لاگ
CLF	اسـ تا ندارد-ه مه منظوره	۱۹۲,۱۶۸,۱,۲۰ - - [۲۶/Dec/۲۰۰۶:۰۳:۰۹:۱۶ ۰۵۰۰] "GET HTTP/ ۱,۱" ۲۰۰ ۱۷۷٤
Apache Custom Log Format	Apache	`LogFormat "%a %l \"%u\" %t %m \"%U\" \"%q\" %p %>s %b %D \"%{Referer}i\" \"%{User-Agent}i\"" my_custom_log` `CustomLog logs/access_log my_custom_log`
W3C Extended Log File	IIS	`#Fields: date time s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-Agent) cs(Referer) sc-status sc-bytes cs-bytes time-taken`

درصورتی‌که فایل‌های دیگر لاگ توسط سرورهای مختلف دیگر تولید شود به‌منظور تحلیل باید به قالب CLF تبدیل شوند. کاربردی‌ترین و مهم‌ترین فیلدها و اجزای اطلاعاتی فرمت CLF در جدول ۵ ارائه‌شده‌اند.

جدول ۵: کاربردی‌ترین فیلدها و اجزای اطلاعات فرمت CLF در لاگ های دسترسی به وب سرور

فیلد داده	توضیحات
IP Address	آدرس IP کاربر درخواست‌کننده
Date_Time Stamp	زمان (تاریخ و ساعت) در قالب [dd/mmm/yyyy:HH:MM:SS -۰۴۰۰] بر اساس تنظیمات منطقه زمانی سرور که درخوا ست کاربر رخ داده است
HTTP Method	شیوه درخواست پروتکل HTTP نظیر GET/POST/HEAD
Hostname	رشته نام میزبان
Requested_URI	مسیر URL درخواست شده کاربر.
HTTP Protocol Version	نسخه پروتکل HTTP که درخواست اجرا شده است
Response Status Code	کدهای وضعیت پاسخ به درخوا ست بر ا ساس شرایط موفقیت‌آمیز بودن ارسال پاسخ به کاربر و سایر شرایط خاص استاندارد

[1] Common Log Format

فیلد داده	توضیحات
Response Size	اندازه شی بازگردانی شده به کاربر به بایت (صرف‌نظر از سرآیندهای HTTP)
Referrer URI	دامنه و مسیر صفحه ارجاع دهنده کاربر با استفاده از سرآیند HTTP
User-Agent	با استفاده از سرآیند HTTP مشخصات دستگاه، سیستم عامل، مرورگر، یا عاملی که کاربر برای درخواست صفحه موردنظر استفاده کرده است را ارائه می‌دهد.

۲-۲-۳-۴ جمع‌آوری لاگ از فایروال‌ها و سیستم‌های امنیت شبکه

فایروال سیستمی است که شبکه را در مقابل نفوذ مهاجمین، دسترسی‌های غیرمجاز، ترافیک‌های مخرب و حملات هکرها محافظت کند. نحوه عملکرد فایروال‌ها به اینگونه است که بسته‌ها را بین شبکه‌ها رد و بدل و مسیریابی می‌کنند. فایروال هم ترافیک ورودی به شبکه و هم ترافیک خروجی از آن را کنترل و مدیریت کرده و با توجه به قوانینی که در آنها تعریف می‌شود به شخص یا کاربر خاصی اجازه ورود و دسترسی به یک سیستم خاص را می‌دهد. بنابراین در اکثر طراحی‌ها تمام ترافیک ورودی و خروجی یک شبکه از فایروال‌های آن می‌گذرد و فایروال‌ها در شبکه‌ها(مانند شبکه اپراتورهای فراهم کننده خدمات اینترنت، اینترانت، داده و غیره) یکی از مناسب‌ترین نقاط برای واکشی داده‌های ترافیکی شبکه هستند و می‌توان توسط آن‌ها ترافیک ورودی و خروجی به شبکه را پایش نمود.

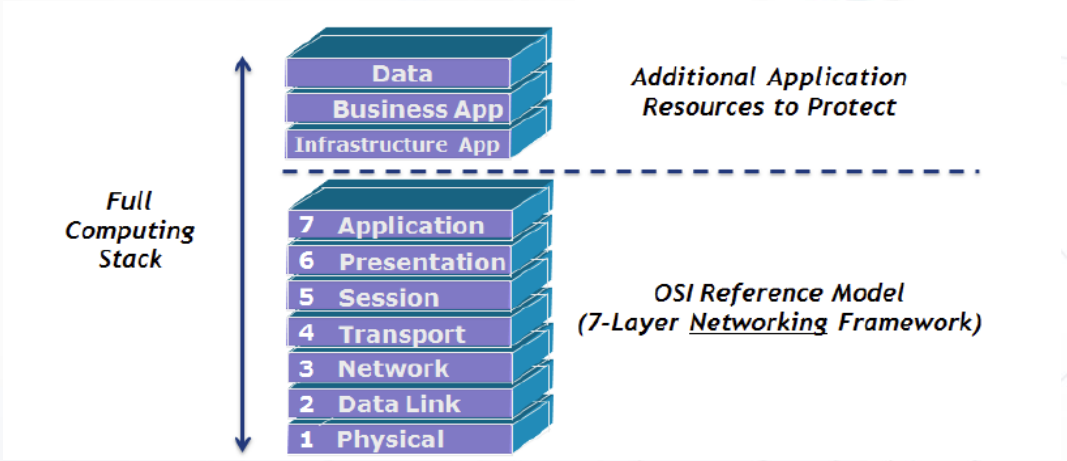
یک سری از فایروال‌ها در سال‌های اخیر بسیار مورد استقبال و کاربرد قرار گرفته‌اند که به نام ForitGate شناخته می‌شوند. با بررسی‌های بیشتر بر روی مستندات و راهنمای این سری از فایروال‌ها مشخص شد که این فایروال‌ها از سیستم‌عاملی معروف به FortiOS بهره می‌گیرد که بر روی آن خدمت/ماژول‌های مختلفی بسته به نیاز و تنظیمات انتخابی کاربر اجرا می‌شوند. از این خدمت‌ها می‌توان به FortiWeb، FortiMail، IPS، UTM و ... اشاره کرد. هرگاه درخواستی به یک فایروال برسد، بسته به اینکه چه ماژول‌هایی روی آن فعال باشند و بسته به اینکه تنظیمات هریک از این ماژول‌ها چه باشد بررسی‌های مختلفی روی آن درخواست انجام می‌شود و Logهای متفاوتی نیز تولید می‌گردد.

به‌عبارت‌دیگر، نوع رکورد لاگ ثبت‌شده و اینکه چه فیلدهایی در هر رکورد لاگ ثبت‌شده باشد بستگی به این دارد که چه ماژول‌هایی فعال هستند، تنظیمات عملکردی و لاگ کردن آن‌ها چیست. به عنوان مثال می‌توان فایروال را طوری تنظیم کرد که تمام ترافیک تبادل شده را لاگ کند، یا اینکه تنها درخواست‌های مرتبط با سایت‌های موجود در black list را ثبت کند یا خطرهای امنیتی احتمالی که IPS تشخیص می‌دهد را لاگ کند. لاگ ترافیک وب توسط فایروال‌ها شامل فیلدهای داده‌ای پروتکل‌های لایه کاربرد مانند HTTP و HTTPS است که در جداول ۱و۲ آورده شده است.

یکی دیگر از خدمت‌های امنیت شبکه، Honey pot هستند. این خدمت یک منبع سیستم اطلاعاتی می‌باشد که بر روی خود اطلاعات کاذب و غیرواقعی دارد و با استفاده از ارزش و اطلاعات کاذب خود سعی می‌کند اطلاعات و فعالیت‌های غیرمجاز و غیرقانونی بر روی شبکه را کشف و جمع‌آوری کند. به زبان ساده Honey pot یک سیستم یا سیستم‌های کامپیوتری متصل به شبکه یا اینترنت است که دارای اطلاعات کاذب بر روی خود می‌باشد و از عمد در شبکه قرار می‌گیرد تا به عنوان یک تله عمل کرده و مورد تهاجم یک هکر یا نفوذگر قرار بگیرد. با استفاده از این

اطلاعات آن‌ها را فریب داده و اطلاعاتی از نحوه ورود آن‌ها به شبکه و اهدافی که در شبکه دنبال می‌کنند جمع‌آوری می‌کند. بنابراین این سیستم یک نسخه از ترافیک ورودی را لاگ می‌کند و با استخراج لاگ این سیستم‌ها می‌توان اطلاعات مفیدی از وضعیت شبکه را تحلیل نمود.

از دیگر خدمت‌های مرتبط با فایروال‌ها و مباحث امنیت شبکه، استخراج لاگ کاربردهای DPI است. فن‌آوری Deep Packet Inspectionیا به اختصار DPI یک نوع فن پردازش و فیلترینگ بلادرنگ (real-time) ترافیک شبکه است که قابلیت دسترسی، تحلیل و اصلاح بخش دیتا یا payload بسته‌های شبکه را امکان‌پذیر می‌سازد. مهم‌ترین خصوصیت این فنّاوری سرعت پردازش بالای حجم بزرگی از داده به‌صورت هم‌زمان است. فن‌آوری‌های پیش از DPI که در اکثر محصولات امنیتی شبکه مثل فایراوال‌ها، IPSها و از این قبیل کاربرد داشتند، تنها توانایی دسترسی به اطلاعات کنترلی یا بخش بسیار اندکی از قسمت داده‌های کاربری را داشتند اما محصولات مجهز به فن‌آوری DPI توسط تکنیک‌های مختلفی قادر به شناسایی، تجزیه‌وتحلیل و درنهایت استخراج کل داده‌های موجود در ترافیک شبکه و تبدیل آن‌ها به اطلاعاتی چون برنامه‌های کاربردی، کاربران، پروتکل‌ها، فایل‌ها، metadataها، انواع ویروس‌ها، بدافزارها و از این قبیل هستند. فن‌آوری DPI این امکان را فراهم می‌سازد تا بتوان بر اساس شاخصه‌های متفاوتی که از بازرسی شبکه استخراج خواهد شد به طبقه‌بندی اطلاعات پرداخت و بر اساس آن توانایی ایجاد و اعمال محدوده وسیع‌تر و دقیق‌تری از سیاست‌های امنیتی و کنترلی را به‌دست آورد[۷].



شکل ۱۲: پشته محاسباتی ترافیک شبکه

می‌توان گفت برای ترافیک وب کشور بیش از ۲۰۰ پروتکل DPI برای ترافیک وب تعریف شده است که این پروتکل‌ها اغلب در دسته‌های زیر می‌گنجد و هریک از حساسیت خاصی در سیاست‌های سایبری، برخوردار هستند و در دست داشتن داده‌های مربوط به ترافیک این پروتکل و پروتکل‌های زیرمجموعه آن به ارائه تحلیل در این دسته‌ها می‌گنجند[۱]:

- پروکسی‌ها که غالباً به‌عنوان فیلترشکن مورد استفاده قرار می‌گیرند.

[۱] مبتنی بر پژوهش‌های پایه‌ای و توسعه‌ای انجام شده

- برنامه‌های کاربردی مانند برخی پیام‌رسان‌ها یا شبکه‌های اجتماعی که کاربران زیادی داشته و ترافیک بالایی را به خود اختصاص داده‌اند و یا به دلایل دیگری حائز اهمیت هستند.
- پروتکل‌های شبکه مانند HTTPS و DNS و FTPS و غیره که احتمالاً به‌طور ترکیبی با قوانین دیگری جهت نیل به اهداف مشخص،در DPI تعریف شده‌اند.
- برخی دامنه‌هایی که به دلایل خاصی حائز اهمیت هستند.

در ادامه لیست برخی از پروتکل‌های DPI مربوط به دامنه‌ها و برنامه‌های کاربردی مختلف آورده شده است:

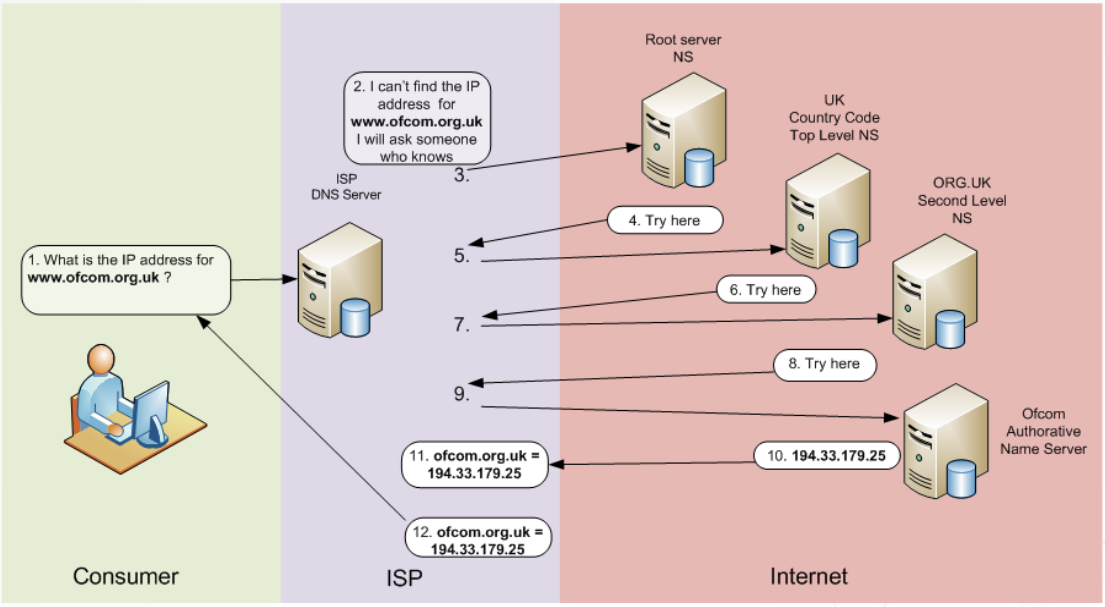
جدول ۶: لیست برخی از کاربردهای DPI

ردیف	نام پروتکل	ردیف	نام پروتکل
۱	All connected	۲۰	Online TV
۲	AlwaysVPN	۲۱	Imo
۳	Anonine	۲۲	Instagram
۴	Ares	۲۳	OpenVPN
۵	AVI	۲۴	Operamini
۶	Baidu	۲۵	Skydrive
۷	BBC iPlayer	۲۶	Skype
۸	BeeTalk	۲۷	PcAnyWhere
۹	Betternet	۲۸	PD proxy
۱۰	Bitmask	۲۹	Persian TV Box
۱۱	BitTorrent	۳۰	PlayStore
۱۲	BlackBerry(tunnel)	۳۱	POP
۱۳	BlackVPN	۳۲	PPP
۱۴	Bleep	۳۳	PPTP
۱۵	Box	۳۴	Private_Internet_Access
۱۶	Browsec	۳۵	Private_Tunnel
۱۷	BufferedVPN	۳۶	Proxifier
۱۸	Chat App	۳۷	ProxPN
۱۹	cisco any connect	۳۸	Proxy Droid

۵-۳-۲-۲ جمع‌آوری لاگ از سرورهای DNS

در یک ارتباط انتها به انتها در شبکه اینترنت و اینترانت از تجهیز کاربر گرفته تا مراکز داده و اپراتورهای فراهم کننده خدمات ارتباطی از پروتکل DNS جهت تبدیل نام دامنه به آدرس IP و مسیریابی پیام‌ها استفاده می‌کنند. بنابراین سرورهای DNS شامل اطلاعات مفیدی برای تحلیل ترافیک این شبکه‌ها هستند. سیستم نام دامنه یا DNS یک پایگاه داده توزیع‌شده از اطلاعات نام‌های دامنه و آدرس‌های IP متناظر آن‌ها است. ماشین‌های موجود در اینترنت بر اساس سیستم آدرس‌دهی IP شناسایی می‌شوند. DNS این امکان را فراهم می‌آورد که بتوان به‌جای آدرس IP ماشین‌ها از طریق نام دامنه به آن‌ها دسترسی داشت. زمانی که نام یک وب‌سایت در مرورگر وارد می‌شود، DNS عمل ترجمه نام دامنه وب‌سایت به آدرس IP منحصربه‌فرد ماشین میزبان را انجام می‌دهد.

نحوه کار سیستم نام دامنه برای پیدا کردن آدرس IP متناظر با یک نام دامنه در شکل ۱۳ نمایش داده شده است.



شکل ۱۳: فرآیند تبدیل نام به آدرس IP

هنگامی که کاربر درخواست مشاهده وب‌سایت www.ofcom.org.uk را در مرورگر تایپ می‌کند، ابتدا یک درخواست DNS برای پیدا کردن آدرس IP ماشین میزبان آن دامنه به سرور DNS ارسال می‌شود. سرور DNS درخواست ترجمه نام را از ماشین کاربر(در اصطلاح resolver) دریافت کرده و فرآیند ترجمه نام دامنه به آدرس IP را انجام می‌دهد. یک بسته درخواست DNS شامل یک RR نوع A است که فیلد نام آن معادل نام دامنه درخواستی و فیلد A آن خالی است. سرور در پاسخ به این درخواست فیلد A را با مقدار IP پر می‌کند و به ماشین کاربر می‌گرداند. اگر سرور نام دامنه محلی، اطلاعات آدرس IP آن مربوط به آن دامنه را نداشته باشد، از دیگر سرورهای نام دامنه، به‌وسیله پرس و جوی تکراری(iterative) آن را به دست می‌آورد[۸].

ازآنجاکه ترافیک DNS رمز شده نیست و بسته‌های جستار به‌صورت متن آشکار بر روی شبکه ارسال می‌شود، دربردارنده اطلاعات نام دامنه و آدرس آی‌پی میزبان‌های آن‌ها است. این اطلاعات می‌تواند به شناسایی تعداد بازدیدهای انجام‌شده از یک وب‌سایت و نیز ردگیری افرادی که یک دامنه را بازدید کردند کمک کند.

جدول ۷، فیلدهای ترافیک قابل جمع‌آوری از درخواست‌های DNS که در لاگ شبکه ثبت می‌شود را ارائه کرده است.

جدول ۷ : فیلدهای ترافیک درخواست‌های DNS در لاگ شبکه

ردیف	نام فیلد	توضیح
۱.	total_len	حداکثر طول فریم(واحد داده در لایه پیوند داده/اترنت) برحسب بایت به همراه سرآیند و پس آیند (MTU+header+trailer)[۱]
۲.	Table_id	N/A
۳.	cdr_time	زمان دریافت رکورد داده تماس
۴.	start_time	زمان شروع ارتباط(درخواست) بر روی تجهیز

[۱] Catalyst ۴۵۰۰/۴۰۰۰ with Supervisor V

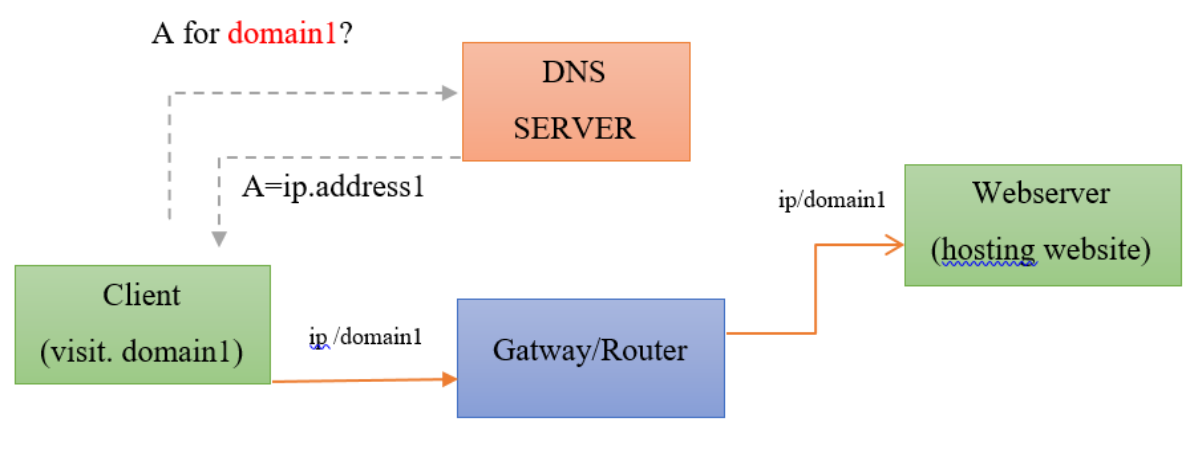
ردیف	نام فیلد	توضیح
۵.	cdr_id	شناسه رکورد داده تماس
۶.	trans_time	زمان سپری‌شده بین شروع و پایان ارتباط. زمان بین شروع ارسال اولین بیت و اتمام آخرین بیت اطلاعات.
۷.	dest_ip	آدرس IP مقصد را مشخص می‌کند.
۸.	user_port	شماره درگاه مبدأ ارتباط را مشخص می‌کند.
۹.	dest_port	شماره درگاه مقصد ارتباط را مشخص می‌کند.
۱۰.	Tcp_fin_pkts	نشان‌دهنده‌های تعداد بسته‌های FIN دریافت شده
۱۱.	up_tcp_dropped_pkts	تعداد بسته‌های TCP ارسالی ازدست‌رفته
۱۲.	down_tcp_dropped_pkts	تعداد بسته‌های TCP دریافتی ازدست‌رفته
۱۳.	up_packets	تعداد بسته ارسالی را تعیین می‌نماید.
۱۴.	down_packets	تعداد بسته دریافتی را تعیین می‌نماید.
۱۵.	up_bytes	حجم داده ارسالی(به بایت) را مشخص می‌کند.
۱۶.	down_bytes	حجم داده دریافتی(به بایت) را مشخص می‌کند.
۱۷.	tcp_window_size	اندازه پنجره زمانی TCP
۱۸.	tcp_mss_size	N/A
۱۹.	tcp_reset_direction	N/A
۲۰.	proto_type	نوع پروتکل (به دلیل حجم لاگ ها قابل تشخیص نیست)
۲۱.	segment_flag	N/A
۲۲.	tcp_syn_pkts	تعداد بسته‌های SYN
۲۳.	tcp_conn_succ	Flag مشخص‌کننده موفقیت‌آمیز بودن connection
۲۴.	tcp_syn_ack_pkts	تعداد بسته‌های SYN ACK
۲۵.	tcp_ack_pkts	تعداد بسته‌های ACK
۲۶.	up_ip_frag_pkts	N/A
۲۷.	down_ip_frag_pkts	N/A
۲۸.	up_tcp_out_order_pkts	تعداد بسته‌های ارسالی خارج از نوبت را نشان دهد.
۲۹.	down_tcp_out_order_pkts	تعداد بسته‌های دریافتی خارج از نوبت را نشان دهد.
۳۰.	up_tcp_retrans_pkts	تعداد بسته‌های ارسالی تکراری
۳۱.	down_tcp_retrans_pkts	تعداد بسته‌های دریافتی تکراری
۳۲.	tcp_reset_pkts	تعداد بسته‌های RST
۳۳.	Direction	N/A
۳۴.	protocol_type	نوع پروتکل مورداستفاده در ارتباط را مشخص می‌کند.
۳۵.	syn_synack_delay	تأخیر بین بسته syn و synack در زمان برقراری ارتباط(مربوط به‌طرف فرستنده در یک ارتباط) را مشخص می‌کند.
۳۶.	synack_ack_delay	تأخیر بین بسته synack و ack در زمان برقراری ارتباط(مربوط به‌طرف گیرنده در یک ارتباط) را مشخص می‌کند.
۳۷.	ack_first_data_delay	تأخیر دریافت ACK مربوط به اولین داده ارسالی را مشخص می‌کند.

ردیف	نام فیلد	توضیح
۳۸.	first_second_data_delay	تأخیر زمانی بین اولین و دومین داده دریافتی مربوط به یک اتصال را مشخص می‌کند.
۳۹.	trans_id	N/A
۴۰.	question_rec	اطلاعات رکورد درخواست در DNS را مشخص می‌کند.
۴۱.	answer_rec	اطلاعات رکورد پاسخ در DNS را مشخص می‌کند.
۴۲.	authority_rec	اطلاعات سرور ارائه‌دهنده اطلاعات پاسخ به درخواست DNS را مشخص می‌کند.
۴۳.	addition_rec	نشانگر اطلاعات اضافی در رکورد پاسخ DNS می‌باشد.
۴۴.	query_type	مشخص‌کننده نوع درخواست DNS ارسالی
۴۵.	query_class	مشخص‌کننده کلاس درخواست DNS ارسالی(معمولاً IN)
۴۶.	resp_delay	مدت‌زمان تأخیر پاسخ در درخواست‌های DNS را مشخص می‌کند.
۴۷.	req_count	تعداد درخواست‌های DNS را نشان می‌دهد.

برای تحلیل روش نگاشت آدرس‌های IP موجود در فایل‌های لاگ HTTPS به رکوردهای متناظر در لاگ DNS از یک سناریو پیاده‌سازی با استفاده از ابزار تحلیل شبکه Wireshark استفاده‌شده است. تئوری در نظر گرفته شده برای این سناریو با این فرض است که ترافیک HTTPS شامل درخواست‌های کاربر به وب‌سایت در دسترس است و ترافیک HTTPS فیلدهای ip address مقصد و شماره درگاه مقصد را دارا است. مشکل آن است که به دلیل آنکه درخواست‌های HTTPS رمز شده است و محتوای بسته‌ها ازجمله فیلدهای URI، Content_type و غیره که در ترافیک عادی HTTP قابل رؤیت بود، قابل مشاهده نیست. در این حالت به دلیل وجود میزبان‌های اشتراکی، به‌صورت معمول نمی‌توان از آدرس IP میزبان در لاگ شبکه دریافت که چه دامنه‌ای در ip/port ماشین میزبان توسط کاربر بازدید شده است. برای حل این موضوع فرضیات زیر در نظر گرفته شده است:

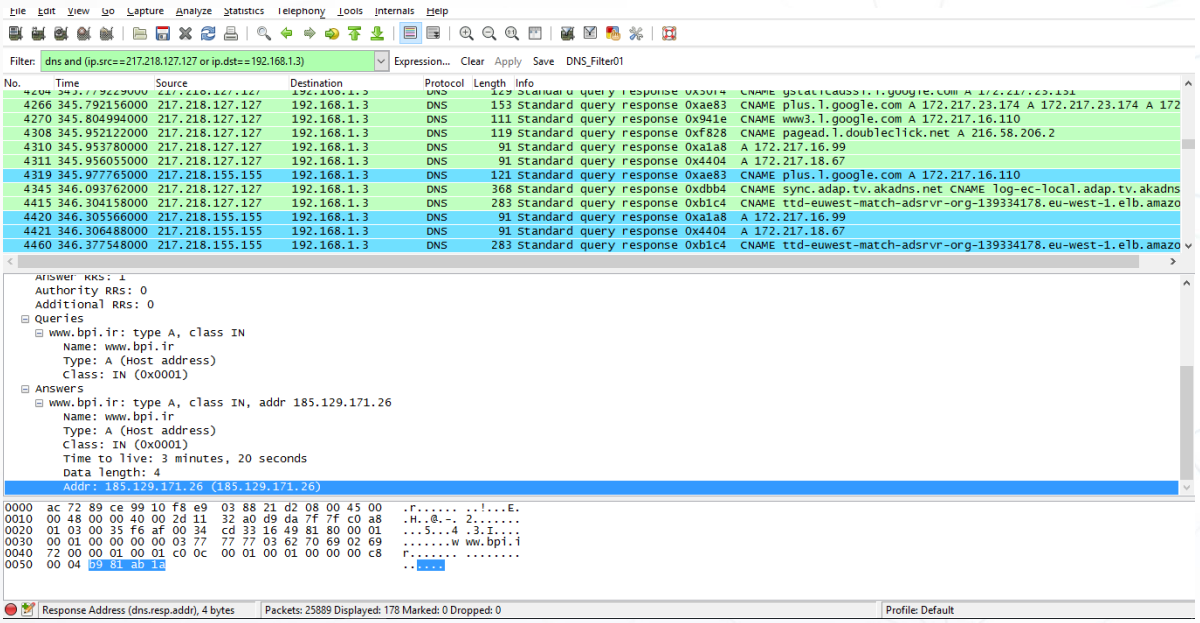
- ترافیک DNS حاوی درخواست‌های ترجمه نام دامنه به آدرس آی‌پی است (پاسخ‌ها نیز قابل مشاهده است)
- ترافیک HTTPS/tls حاوی آدرس ip مقصد و درگاه وب سروری است که سایت موردنظر برای بازدید را میزبانی می‌کند.
- درصورتی‌که ip آدرس مقصد درخواست‌ها و زمان درخواست‌ها در HTTPS در دسترس باشد، برای مشخص شدن دامنه بازدید شده توسط کاربر می‌توان در محدوده زمانی مشابه به درخواست‌های ارسال شده به DNS مراجعه کرد و تعیین نمود دامنه مورد بازدید کاربر بر روی آن میزبان کدام دامنه بوده است.

شکل ۱۴ الگوی مشاهده یک صفحه وب توسط کاربر را نمایش داده شده است.



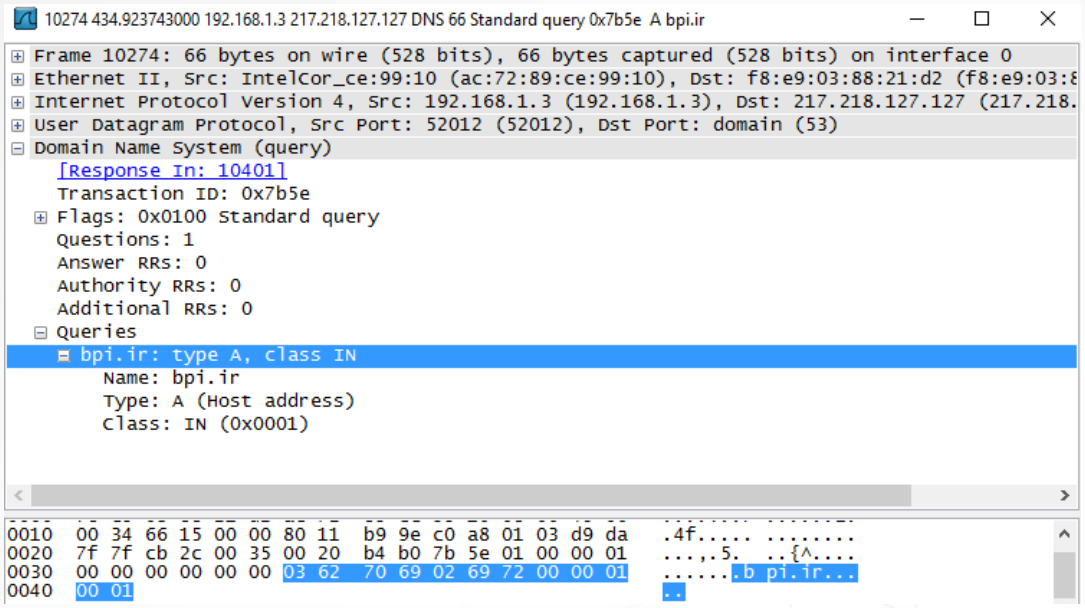
شکل ۱۴: مشاهده صفحه وب توسط کاربر

سناریوی مطرح شده در بالا بر روی ترافیک واقعی و یک وب‌سایت نمونه موردبررسی قرار داده شده است، هدف اصلی این کار استفاده از تحلیل بسته‌های DNS و شناسایی فیلدها به‌منظور دستیابی به نام‌های دامنه متناظر در هر بازدید از میزبان HTTPS است.



شکل ۱۵: مشاهده ترافیک DNS برای بازدید از دامنه HTTPS://bpi.ir

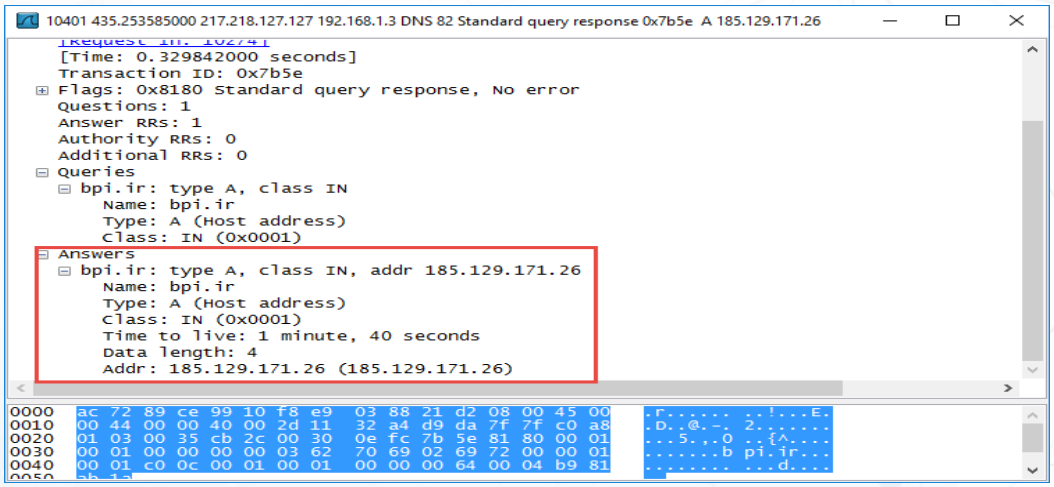
همان‌طور که مشاهده می‌شود، ترافیک ضبط‌شده از بسته‌های پرس و جوی DNS به دامنه www.bpi.ir که در پاسخ آدرس IP برگردانده شده است، حاوی نام دامنه درخواستی و آدرس پاسخ داده شده از سمت سرور نام دامنه است. جزئیات اطلاعات مربوط به این بسته‌ها در شکل ۱۶ نمایش داده شده است



شکل ۱۶: جزئیات کوئری درخواست dns

مشاهده می‌شود که درخواست ارسالی حاوی نام دامنه موردنظر کاربر است. در این مرحله رکورد A که بیانگر آدرس دامنه است خالی است.

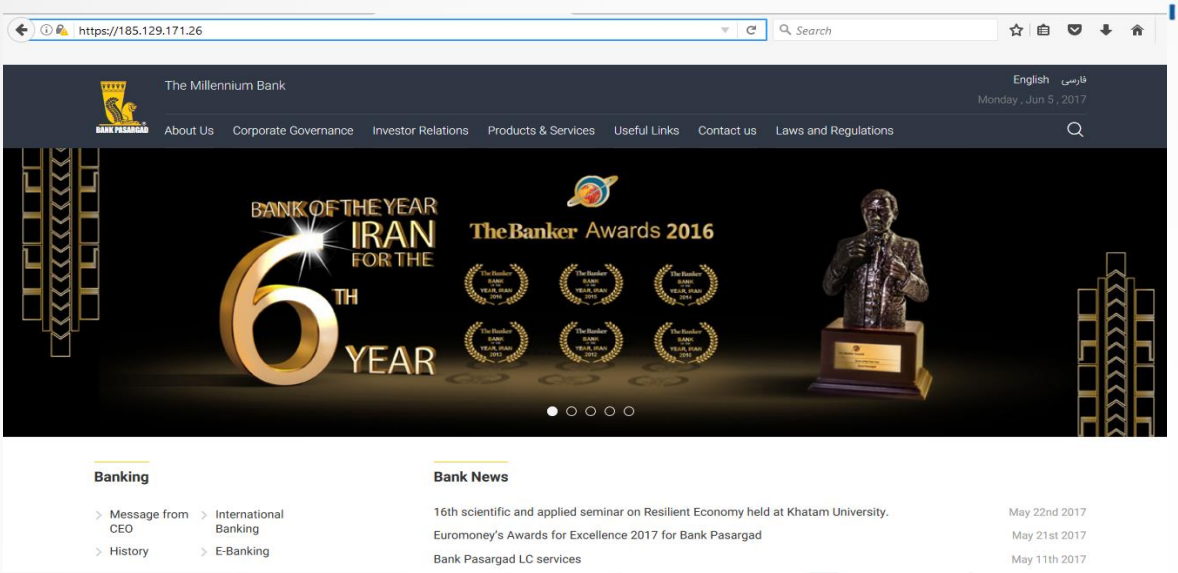
در ادامه جزئیات بسته پاسخ دریافتی که حاوی آدرس ip وب‌سایت مورد درخواست کاربر است در ترافیک قابل دستیابی است. شکل ۱۷ رکورد متناظر پاسخ سرور DNS را نشان می‌دهد.



شکل ۱۷: جزئیات پاسخ dns

سرور پاسخ درخواست ترجمه نام به آدرس ip را داده است و آدرس ip میزبان وب‌سایت را در رکورد A بازگردانده است. همچنین با درج آدرس IP به‌دست‌آمده از ترافیک DNS در مرورگر صفحه انگلیسی زبان سایت بانک پاسارگاد نشان داده می‌شود (شکل ۱۸).

HTTPS://۱۸۵,۱۲۹,۱۷۱,۲۶



شکل ۱۸: دسترسی به وب‌سایت با استفاده از آدرس ip میزبان (ip اختصاصی)

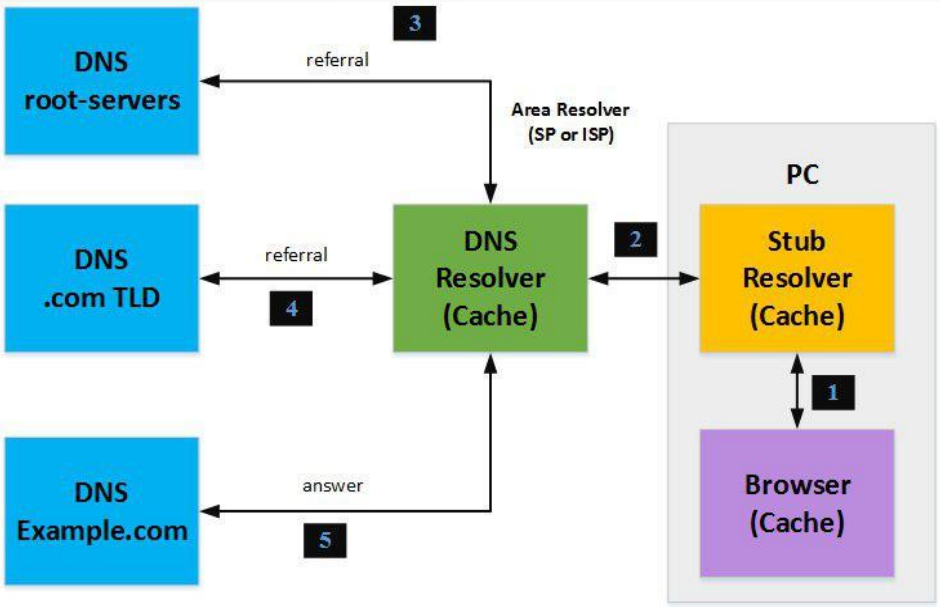
این امر که از مسیر بتوان از آدرس IP یک سایت به آن سایت دسترسی داشت در همه موارد امکان پذیر نیست. علت این امر استفاده از میزبان‌های اشتراکی دراینترنت است. میزبان‌های اشتراکی این قابلیت را ارائه می‌دهند که بتوان چندین نام دامنه را بر روی یک آدرس IP میزبانی کرد. بنابراین با داشتن آدرس IP به تنهایی، نمی‌توان به یک دامنه مشخص در میزبان اشتراکی دسترسی داشت.

حافظه نهان[۱] DNS یک پایگاه داده موقت در کامپیوتر است که حاوی اطلاعات مربوط به دامنه‌های مورد بازدیدی است که کاربر اخیرا مشاهده کرده است. به‌عبارت‌دیگر کش حافظه ای از جستارهای اخیر DNS نگهداری می‌کند. هنگامی که اطلاعات مربوط یک پرس و جوی DNS در کش موجود باشد در هر بار بازدید به جای ارسال درخواست بر روی درگاه DNS در شبکه از اطلاعات موجود در کش محلی استفاده می‌شود. به‌طورکلی پرس و جو های DNS در سه مرحله زیر ممکن است کش شوند:

- حافظه نهان مرورگر
- حافظه نهان سیستم‌عامل (DNS Resolver cache)
- حافظه نهان سرور نام دامنه (DNS Server cache)

شکل ۱۹ مکان‌های ممکن برای پرس و جوهای DNS را نشان می‌دهد. بسیاری از اپراتورهای فراهم کننده خدمت اینترنت و مراکز داده جهت بالا بردن سرعت تبدیل IP سرورهای حافظه نهان DNS در شبکه خود مستقرکرده‌اند و درخواست کاربر قبل از اتصال به سرورهای DNS در سلسله مراتب بالاتر، به سرور حافظه نهان محلی در شبکه اپراتور ارسال می‌شود و این امر موجب افزایش سرعت دسترسی کاربران به وب‌سایت‌ها می‌شود. بنابراین لاگ این سرورها می‌تواند اطلاعات مفیدی برای تحلیل بازدیدهای کاربران از دامنه‌های وب در اختیار تحلیلگران قرار دهد.

[۱] cache



شکل ۱۹: مکآن‌های ذخیره کش درخواست‌های DNS

۲-۲-۴نمونه‌گیری از داده‌های ترافیک شبکه

جمع‌آوری داده‌ها باید به‌گونه‌ای انجام گیرد که با کل داده‌های شبکه را شامل شود یا توزیع یکنواختی از داده‌ها که به عنوان مثال از نمونه های تصادفی از کل مبادی داده در تمام ساعات شبانه روز تشکیل شده باشد. در مواردی که حجم داده‌ها بسیار بالا است و امکان جمع‌آوری کل داده‌های شبکه وجود ندارد(مانند داده‌های شبکه‌ی زیرساختی کشور) می‌توان از روش‌های نمونه‌گیری استفاده کرد.

نمونه گیری یک روش معمول برای انتخاب یک زیر مجموعه از داده‌ها برای تحلیل داده است. نکته کلیدی نمونه گیری این است که باید از نمونه ای استفاده کنیم که نماینده کل داده باشد چنین نمونه ای باید تقریبا به اندازه داده اصلی ویژگی داشته باشد و اگر میانگین ویژگی داده اصلی برابر مقدار خاصی باشد میانگین داده نماینده نیز باید نزدیک به همان عدد باشد. انواع نمونه‌گیری شامل موارد زیر است[۹]:

- نمونه گیری تصادفی عادی[۱] : شانس برابری برای انتخاب هر مورد از داده‌ها وجود دارد.
- نمونه گیری بدون جایگزین[۲]: هر نمونه که انتخاب می‌شود از جامعه حذف می‌شود.
- نمونه گیری با جایگزینی[۳]: نمونه‌ها از جامعه پس از انتخاب از جامعه حذف نمی‌شوند.
- نمونه گیری طبقه بندی شده[۴]: داده‌ها به چند دسته تقسیم شده و سپس از هر دسته به صورت تصادفی انتخاب می‌شود.

[۱] Simple Random Sampling
[۲] Sampling without replacement
[۳] Sampling with replacement
[۴] Stratified sampling

- نمونه گیری تصاعدی[۱] :پیدا کردن اندازه نمونه مناسب می‌تواند سخت باشد در نتیجه طرح‌های انطباقی و تصاعدی در بعضی موارد استفاده می‌شود. این روش‌ها از یک نمونه کوچک شروع کرده و سپس اندازه نمونه‌ها را افزایش داده تا به یک اندازه مناسب دست پیدا کنند.

[۱] Progressive Sampling

۳ تحلیل ترافیک شبکه

تجزیه و تحلیل فرآیندی علمی جهت تبدیل داده به بینش و تصمیم گیری بهتر و متعاقب آن اقدامات مناسب است. دو فرآیند کلی را می توان برای تجزیه و تحلیل در نظر گرفت، ۱) تولید بینش از داده و ۲) تصمیم گیری بر اساس بینش تولید شده . شکل ۲۰ فرآیند تجزیه و تحلیل را نشان می دهد.

شکل ۲۰: فرآیند تجزیه و تحلیل داده

همانگونه که در شکل ۲۰ نشان داده شده است، فاصله بین داده تا بینش با تجزیه و تحلیل تکمیل می گردد و به نوعی می‌توان گفت که این موتور تجزیه و تحلیل است که با دریافت داده، بینش لازم را تولید می‌کند. همچنین، فاصله بین زمانی که به بینش لازم رسیده تا اقدام متناسب با آن انجام شود، به تصمیم گیری برمی گردد. سرعت، دقت و کیفیت در هر دو مرحله بسیار تعیین کننده و مهم می باشد چراکه زندگی امروز در یک فضای شدیدا رقابتی شکل گرفته است. چه بسیار سازمان هایی که علی‌رغم تولید بینش لازم از داده، درمرحله تصمیم گیری دارای اشکالات فراوان هستند و به دلیل نبود یک ساختار منسجم تصمیم گیری، تصمیمات با تاخیر و احتمالا با ابهام اتخاذ می‌کنند. همین موضوع باعث می شود که آن‌ها نتوانند در زمان لازم اقدامات مناسب و بایسته ای را انجام دهند. شاید اگر امکان داشت تا نقش انسان را در فرآیند تصمیم‌گیری کمتر کرد، فاصله زمانی بین داده تا اقدام نیز کمتر می‌شد.

چه میزان از فرآیند تصمیم گیری خودکار است و چه میزان از آن با دخالت عوامل انسانی انجام می شود ؟ پاسخ به این پرسش می‌تواند علل ایجاد سطوح مختلف تجزیه و تحلیل را تشریح نماید. در واقع، می توان سطوح مختلفی از بلوغ تجزیه و تحلیل را با توجه به میزان خودکار بودن تصمیم گیری تعریف نمود . شکل ۲۱ سطوح مختلف بلوغ تصمیم گیری را نشان می‌دهد[۱۰].

شکل ۲۱: سطوح مختلف بلوغ تصمیم‌گیری

در شکل ۲۱ سوالاتی نظیر چه اتفاقی افتاده است ؟ چرا این اتفاق افتاده است ؟ چه اتفاقی خواهد افتاد و چه کاری را می بایست انجام داد به خوبی تعریف شده اند . تفکیک آن بخشی از تجزیه و تحلیل که به صورت اتوماتیک انجام می شود و آن بخشی که توسط انسان انجام خواهد شد، یکی از نکات مهم مدل نشان داده شده در شکل فوق می باشد.

هر نوع و یا مدل تجزیه و تحلیل نشان داده شده در شکل فوق، از روش ها و الگوریتم‌های مختلفی استفاده می‌کنند. این بدان معنی است که با توجه به نوع نیاز و انتظار یک سازمان از تجزیه و تحلیل، تنوع در بکارگیری نوع های مختلف داده، امکانات ذخیره سازی و پردازش وجود دارد تا نتایج دلخواه را ارایه نماید. تجزیه و تحلیل توصیفی بر روی داده‌ها معمولا به سه روش انسانی(خبرگان و کارشناسان حوزه تحلیل)، استفاده از فناوری هوش مصنوعی و یا روش‌های آماری و ریاضی انجام می‌پذیرد. تحلیل توصیفی روی داده‌های ترافیک شبکه نیز از این حالات مستثنی نیست. در این روش‌ها داده‌ها پس از پردازش، اطلاعاتی را تولید می‌کنند که این اطلاعات به مسئولان مرتبط جهت مدیریت بهتر و اتخاذ تصمیم‌ها و سیاست‌گذاری‌ها ارائه می‌شود. در ادامه به معرفی روش‌هایی برای تحلیل داده‌های ترافیک شبکه پرداخته شده است.

۳-۱ پایش و مدیریت شبکه با ارائه تحلیل‌های توصیفی بر روی لاگ ترافیک شبکه

در تجزیه و تحلیل توصیفی، هدف اصلی پاسخ به این نوع سوالات است:

- چه اتفاقی افتاده است ؟
- چه چیزی دارد اتفاق می‌افتد؟
- چه چیزی ممکن است اتفاق بیافتد؟

تا در نهایت برای تصمیم گیری بینش تولید شده را پیش روی تصمیم گیرندگان انسانی قرار دهد. در این نوع تجزیه و تحلیل با دستیابی به سوابق که ریشه در گذشته و حال دارد، امکان تحلیل داده مبتنی بر رویدادهای گذشته و حال و تولید بینش مسیر پیش رو، میسر می‌گردد. امکاناتی نظیر BI، OLAP، داشبوردها، گزارشات و بصری سازی را می‌توان در این گروه قرار داد. گوگل آنالیتیکز[۱] یک نمونه شناخته شده در ارائه تحلیل توصیفی وب است. با تجزیه و تحلیل توصیفی، می‌توان دنیا را به گونه‌ای که هست توصیف کرد .

از منظر روش تجزیه و تحلیل و میزان خودکارسازی فرآیند انجام کار، می‌توان تحلیل‌های توصیفی را به دو قسم تحلیل‌های مبتنی بر هوش مصنوعی و تحلیل‌های آماری تقسیم نمود. **هوش مصنوعی** علم ساخت ماشین های است که مانند انسان فکر می‌کنند، مانند انسان عمل می‌کنند، منطقی فکر می‌کنند و عمل می‌کنند. قابلیت‌های مورد نیاز برای انسان گونه عمل کردن، اقداماتی همچون پردازش زبان طبیعی، ذخیره و بازیابی دانش، استدلال خودکار و یادگیری عمیق یا یادگیری ماشین است. درک مفاهیم تحلیل شبکه‌ها مستلزم شناخت شبکه‌ها و ابزارها و فناوری‌های مختلف مورد نیاز برای تحلیل آن‌ها است که هوش مصنوعی مهم‌ترین فناوری مورد استفاده برای تحلیل شبکه‌های اجتماعی است.

هدف از **تحلیل‌های آماری**، کشف و نمایش الگوهای ترافیکی برگرفته از تحلیل آماری رفتار شبکه و کاربران است. تحلیل‌های آماری روی داده‌های خام و یا اطلاعات حاصل از سایر تحلیل‌های توصیفی و یا به صورت ترکیبی از آن‌ها

[۱] Google Analytics

انجام می‌شود. به عنوان مثال تحلیل آماری روی اطلاعات دسته‌بندی دامنه‌های وب و همچنین داده‌های بازدید از آن‌ها می‌تواند جهت‌گیری و علایق افراد در بازدید از دامنه‌های وب در آن شبکه را نشان دهد.

با دیدگاه مدیریت شبکه، تحلیل آماری ترافیک شبکه در لایه کاربرد ابزاری قدرتمند جهت استخراج الگوهای رفتاری کاربران در شبکه است و امکان رصد عمیق‌تر و دقیق‌تر جریانات آن شبکه را در اختیار مسئولان و سیاست‌گذاران آن شبکه می‌گذارد.

تحلیل‌های آماری و مبتنی بر هوش مصنوعی متنوعی می‌تواند به پایش ترافیک شبکه (در راستای مدیریت بهتر شبکه و سیاست‌گذاری برای آن) کمک کند؛ نمونه‌ای از این تحلیل‌ها عبارتند از:

- تحلیل بازدید کاربران از وب‌سایت‌ها (تحلیل وب) شامل انواع تحلیل‌های آماری و هوشمند(دسته‌بندی، تحلیل احساسات، تحلیل تعاملات، تحلیل شبکه‌های اجتماعی و ...)
- تحلیل دسته‌بندی ترافیک بر اساس معیارهای مشخص مانند دسته‌بندی موضوعی کاربردهای شبکه
- تحلیل احساسات روی داده‌های کیفیت خدمت و میزان رضایتمندی کاربران شبکه
- تحلیل‌های مبتنی بر گراف به منظور شناسایی الگوی ترافیک و کمک به شناخت وضعیت و باگ‌های شبکه
- پایش وضع موجود و ارائه تحلیل‌های پیش‌بینی وضعیت آینده شبکه به منظور ایجاد سیستم‌های یکپارچه مدیریت شبکه، برای جلوگیری از وقوع مشکلات، خرابی‌ها، مدیریت شرایط بحرانی و ازدحام ترافیک (Congestion)
- تحلیل میزان بهره‌وری کل شبکه با استفاده از داده‌های آموزشی
- تحلیل روندهای تغییر الگوهای ترافیک مشتریان و وفق پذیری با آن
- تحلیل های پایداری شبکه و نظایر آن

شرح راه‌کارها و ارائه الگوریتم‌های هوش مصنوعی، یا تعریف روش‌ها و نمونه‌های آماری به ازای تمام موارد تحلیلی نامبرده فوق و سایر موارد مشابه به دلیل حجم بالای کار و همچنین نیاز به شناخت دقیق داده‌های قابل گردآوری و سیاست‌ها و نیازمندی‌های تحلیلی شبکه ملی اطلاعات در این گزارش آورده نشده است و می‌تواند در قالب یک پروژه تعریف شود؛ با این حال در ادامه به عنوان نمونه و جهت درک بهتر ماهیت و عمق مطالب، شرح بیشتری از تحلیل توصیفی هوشمند دسته‌بندی دامنه‌های وب بازدید شده در ترافیک شبکه و همچنین تحلیل‌های آماری روی ترافیک بازدید کاربران از دامنه‌های وب در شبکه ملی اطلاعات ارائه شده است.

۳-۱-۱ تحلیل دسته بندی محتوای متنی وب‌سایت‌های بازدید شده توسط کاربران شبکه

به منظور دسته‌بندی محتوای متنی تحت وب، ابتدا باید موضوع مربوط به هرکدام از این محتواهای متنی تشخیص داده شده و سپس بر اساس این موضوعات دسته‌بندی آن‌ها انجام شود. انجام تحلیل‌های مختلف مانند تحلیل دسته‌بندی دامنه‌های وب نیاز به داشتن محتوای هریک از دامنه‌های وب بازدید شده در لاگ ترافیک شبکه است. یکی از رویکردهای اصلی در حوزه گردآوری داده از وب، خزش وب[۱] می‌باشد. یک خزشگر وب یک اسکریپت

[۱] web crawling / web scrapping

نرم افزاری یا برنامه ریزی است که وب سایت‌ها را به شیوه‌ای منظم و خودکار مرور می‌کند. معمولاً خزش‌گرها با کاربرد شناسایی تمام دامنه‌ها و لینک‌های موجود در آن برای موتورهای جستجو مورد استفاده قرار می‌گیرند.

پژوهش‌ها و توسعه‌های انجام شده بر روی مجموعه‌ای از دامنه‌های بازدید شده توسط کاربران که در لاگ شبکه وجود داشته‌اند، نشان می‌دهد که به منظور دسته‌بندی و تشخیص موضوع، از **الگوریتم‌های یادگیری ماشین**[۱]، **یادگیری عمیق**[۲] و **قوانین**[۳] استفاده می‌شود.

قبل از انجام هر گونه تحلیلی بر روی محتوای متنی جمع‌آوری شده، نیاز است تا اعمال فرآیند استخراج ویژگی (در صورت بکارگیری الگوریتم یادگیری ماشین برای دسته‌بندی) به‌منظور ساخت بردار ویژگی از متنِ مورد بحث و آماده‌سازی آن برای اعمال الگوریتم یادگیری ماشین انجام پذیرد. همچنین قبل از انجام هر تحلیلی در همه متون و از جمله متن‌های موجود در وب‌سایت‌ها، لازم است پیش‌پردازش‌هایی در این متن‌ها صورت گیرد که متن را به حالت استاندارد و نرمال تبدیل کرده تا برای تحلیل‌ها و پردازش‌های بعدی آماده باشند. هدف از این پیش پردازش‌ها تسهیل فرایند درک متن توسط کامپیوتر است. مهم‌ترین تکنیک‌های پیش پردازش داده‌ها عبارتند از:

- تکنیک‌های پاکسازی داده [۴]، هدف آنها از بین بردن داده‌های نویزی و ناسازگاری‌های بین داده‌ها است.
- تکنیک‌های یکپارچه‌سازی داده[۵]، از آنجایی که ممکن است داده‌ها از منابع مختلفی جمع آوری شده باشند، نیاز به یکپارچگی بین آنها است.
- تکنیک‌های کاهش داده[۶]، در حجم بالای داده‌ها ممکن است بعضی از داده‌های غیر مفید هم وجود داشته باشند و نیاز نباشد همه داده‌ها در پردازش نهایی باشند، این تکنیک‌ها در این موارد کاربرد دارند.
- تکنیک‌های تبدیل داده‌ها[۷]، این تکنیک‌ها هنگامی مفید خواهند بود که لازم باشد بر روی داده‌ها نرمال‌سازی‌هایی انجام شود.

برای دسته‌بندی دامنه‌ها به‌طورکلی می‌توان آن‌ها را به دو دسته تقسیم نمود. دامنه‌هایی که دارای محتوای متنی هستند و دامنه‌هایی که بدون محتوای متنی هستند. ازاین‌رو در صورت وجود محتوای متنی قابل‌اتکا می‌توان توسط الگوریتم‌های یادگیری ماشین با استفاده از روش های با نظارت و از روش‌های مبتنی بر قوانین برای دامنه‌های فاقد محتوا، دسته‌بندی‌ها را انجام داد.

۳-۱-۱-۱ الگوریتم‌های یادگیری ماشین بانظارت با استفاده از خزش محتوای آدرس‌های ULR مورد بازدید کاربران

در این رویکرد کل محتوای متنی موجود در صفحه، اعم از متون ساده و متون دارای لینک به عنوان متن ساده در نظر گرفته می‌شود و کل متن به صورت یکپارچه تحلیل می‌شود. جهت افزایش دقت دسته‌بندی بهتر است تا سیاست خزش با عمق بیشتری از صفحه[۸]، مدنظر قرار گیرد. یکی از کاربردهای متن‌کاوی در سیستم‌های پاسخگویی

---

[۱] Machine learning

[۲] Deep learning

[۳] Rule based algorithems

[۴] Data cleaning

[۵] Data integration

[۶] Data reduction

[۷] Data transformations

[۸] منظور از عمق صفحه، خزش لینکهای یک صفحه و به همین ترتیب لینکهای لینکهای صفحه و ...

به سوالات[۱] روش‌های یادگیری ماشین با نظارت می‌باشد که این سیستم‌ها از مهم‌ترین برنامه‌های کاربردی در سیستم‌های بازیابی اطلاعات هستند. در همین راستا و به منظور بهینه‌سازی سیستم‌های پاسخگویی به سوالات، به ارائه یک فریم‌ورک مبتنی بر گرامر[۲] به منظور دسته‌بندی سوالات کاربران پرداخته است[۱۱]. در ابتدا مجموعه‌ای از سوالات مختلف در راستای مشخص نمودن انواع موضوعات بازدید شده توسط کاربران از خبرگان تحلیل لاگ شبکه پرسیده، جمع‌آوری و مورد تحلیل واقع شده است. این مراحله تعیین برچسب‌ها نام دارد. سایر مراحل انجام کار در سه فاز اصلی توسط ماشین انجام می‌شود؛ در فاز اول نیاز است تا بر اساس دو منبع مختلف که یکی قواعد و ساختار گرامری و دیگری نوع اطلاعات ارائه شده در سایت مذکور[۳] می‌باشد، یک term taxonomy[۴] یالغتنامه طراحی و ساخته شود که در این لغتنامه به‌گونه‌ای معنا و ساختار همه لغاتِ موجود در سوالات پوشش داده شود. در فاز دوم نیز به ازای هر برچسب بر اساس لغتنامه تعیین شده برای آن، درخت پارس به طور مجزا ترسیم شده و درواقع سؤال مذکور به یک درخت نگاشت می‌شود. در فاز سوم هدف طراحی دسته‌بند[۵] برای دسته‌بندی دامنه‌ها بر اساس درخت‌های پارس ایجاد شده در مرحله قبل می‌باشد. بنابراین تعدادی از دامنه‌ها به منظور ساخت و مدل کردن ماشین یادگیری (درخت تصمیم) و تعدادی نیز به منظور ارزیابی درخت تصمیم مورد استفاده قرار خواهند گرفت. ازآنجاکه این فریم‌ورک پیشنهادی مبتنی بر قواعد و ساختار گرامری زبان طراحی شده است، مهم‌ترین مزیت آن، قابلیت تعمیم به زبان‌های مختلف (البته به شرط فراهم بودن ابزارها و پیمانه‌های گرامری آن زبان) است. بنابراین قبل از هرچیز در فرآیند دسته‌بندی متون، یکی از مهم‌ترین مراحل، استخراجِ بهینه ویژگی‌های مربوطه از متن با استفاده از روش‌های آماری می‌باشد. بدین منظور اسکیماهای شناخته شده‌ای مانند TF و IDF وجود دارد که همگی با تکیه بر رویکرد وزن‌دهی لغات[۶] طراحی شده‌اند. در نهایت ویژگی‌های استخراج شده از متن به مدل با روش‌هایی همچون **خوشه‌بندی** و **الگوریتم چند دسته‌ای** و غیره داده می‌شود تا داده‌ها برچسب‌گذاری شده و برای آن هریک از برچسب‌ها(دسته‌ها) مدل، به طور جداگانه آموزش داده شود و در نهایت با افزایش داده‌های نمونه به دقت کافی در برچسبزنی برسد[۱۲].

با توجه به یک نتیجه عملی و شهودی در بررسی داده‌های ترافیک وب در بازه زمانی بیش از یک سال دامنه‌های بازدید شده در آن ترافیک شبکه به دسته‌ها یا برچسب‌های زیر مدل شدند:

۱. موتورهای جستجو
۲. خدمات ایمیل
۳. اخبار و خبرگزاری
۴. شبکه‌های اجتماعی
۵. پیام‌رسان‌ها
۶. بازی

---

[۱] Question Answering (QA)

[۲] grammar-based framework

[۳] Knowledge Domain

[۴] نوع هر لغت را به لحاظ گرامری و نیز به لحاظ محتوایی مشخص میسازد، مثلاً لغت news به لحاظ گرامری اسم می‌باشد و به لحاظ معنایی جز دسته news and event در سایت قرار می‌گیرد.

[۵] Classifier

[۶] Term Weighting

۷. سایت‌ها و وبلاگ‌های عمومی
۸. اشتراک‌گذاری: امکان اشتراک‌گذاری فایل و یا هر محتوای دیگر را داشته باشند.
۹. آرشیوهای دانلود و پخش محتوای چندرسانه‌ای (مانند تلوبیون-آپارات و ..)
۱۰. آرشیو محتوای نرم‌افزاری (مانند p۳۰download, soft۹۸.ir)
۱۱. فروشگاه‌های آنلاین
۱۲. رزرواسیون مانند eghamat۲۴.ir
۱۳.سایت‌های واسطه فروش مانند divar.ir
۱۴.دامنه‌های ناشناس[۱]

لازم به ذکر است که برچسب‌ها یا دسته‌هایی که توسط مدل‌ها ایجاد می‌شود کاملا مبتنی بر داده‌های آموزشی که در نهایت برگرفته از داده‌های اصلی جاری هستند، می‌باشد و با توجه به کاربردهای مورد استفاده کاربران شبکه متفاوت خواهد بود.

۳-۱-۱-۲ دسته‌بندی مبتنی بر قوانین

در این رویکرد، منابعی جهت دسته‌بندی مدنظر قرار داده می‌شوند که پردازش آن‌ها بر روی آدرسURL صفحه وب و یا آدرس‌های زیردامنه‌های مربوط به آن متمرکز شده است. در روش حل مساله مبتنی بر قوانین، نه مبتنی بر روش‌های یادگیری ماشین و نه صرفاً بر پایه روش‌های ریاضی و آمار بوده است. درواقع روش حل مساله کاملا مبتنی بر یک‌سری قوانین خاص ابداعی از جانب تحلیلگر لاگ ترافیک شبکه است که تقریباً عملکردی مشابه استنتاج در سیستم‌های خبره را خواهد داشت. یعنی ابتدا ایجاد یک پایگاه دانش که قوانین ابداعی در آن غالباً مبتنی بر شرط توسط خبرگان نوشته خواهند شد و سپس ارزیابی داده‌های کاندید که بر اساس این قوانین انجام می‌شود.

روش مبتنی بر قوانین بالاخص برای دسته بندی دامنه‌های وب بدون محتوا و یا دامنه‌هایی که توسط دسته‌بند به‌عنوان دامنه ناشناس دسته‌بندی شده‌اند، کاربرد دارد. به عنوان مثال قوانین زیر بر اساس شواهد روی تعداد بالایی از دامنه‌های نمونه برداری شده در مدت زمان بیش از یک سال قابل استفاده است:

- وجود کلیدواژه‌های api و cloud و dns و map و Push و Pushnotification و update و AS بدون حساسیت در حروف، در نام دامنه نشان می‌دهد که با احتمال بالایی این دامنه در زیر دسته‌های مختلف از دسته خدمت می‌باشد. دسته خدمت به عنوان دسته‌ای تعریف شده است که ترافیک بازدید از دامنه‌ها از طریق تجهیز ارتباطی کاربر در شبکه به طوری که کاربر معمولا از آن آگاه نیست و بدون اراده وی در پشت پرده برنامه‌های کاربردی نصب شده در تجهیز کاربر، انتقال داده می‌شود.
- وجود کلیدواژه‌ی CDN نشان می‌دهد که با احتمال بالایی این دامنه در دسته شبکه‌های توزیع محتوا می‌باشد. این دسته از دامنه‌ها نیز معمولا هنگامی که کاربر درخواست بازدید از یک دامنه با محتوای سنگین مانند، دیجی کالا، دیوار و ... را ارائه می‌دهد به طور ناآگاهانه و جهت دریافت محتوای مورد نیاز کاربر، برای وی ترافیک ایجاد می‌کند.
- وجود کلیدواژه‌هایADو ADVERT و ADVERTISE و ADV و ADVERTISING و ADS بدون حساسیت حروف، در نام دامنه نشان می‌دهد که با احتمال بالایی این دامنه در دسته تبلیغات می‌باشد.

[۱] هر دامنه‌ای که در هیچیک از ۱۳ دامنه فوق جای نگیرد، در این دسته توسط ماشین قرار داده خواهد شد.

شواهد نشان می‌دهد که ترافیک این دسته از دامنه‌ها نیز به طور ناآگاهانه توسط کاربر، زمانی که در حال بازدید از یک برنامه‌کاربردی یا وب‌سایت هستند، به طور ضمنی و از طریق آن وبسایت یا برنامه‌کاربردی ایجاد می‌شود.

روش‌های مبتنی بر قوانین نیز کاملا وابسته به داده‌های ترافیک شبکه در دسترس توسط خبرگان و تحلیلگران لاگ ترافیک شبکه، تعریف می‌شوند و همچنین در ابعاد مختلفی از آن ترافیک می‌توانند پردازش شوند.

۳-۱-۲تحلیل‌های آماری روی ترافیک بازدید کاربران از دامنه‌های وب

به طور کلی تحلیل‌های آماری با نمودارسازی از مجموع، اختلاف، میانگین یا متوسط یک معیار کمی در واحد زمان بر روی یکسری داده یا اطلاعات بدست می‌آید. تحلیل آماری روی ترافیک وب را می‌توان به تحلیل بازدیدهای کاربران و تحلیل حجم ترافیک تبادلی کاربران تقسیم نمود. در ادامه برخی از انواع تحلیل‌های آماری در لایه کاربرد شبکه بر اساس تعریف برخی معیارهای اصلی، آورده شده است.

۳-۱-۲-۱ معیارهای تحلیل آماری ترافیک وب

فیلدهای داده‌ای که از لاگ پروتکل‌های مختلف شبکه قابل جمع‌آوری هستند، در صورتی‌که پردازشاتی روی آن‌ها صورت گیرد، فیلدهای داده‌ای بدست می‌آیند که این فیلدها را می‌توان به عنوان معیارهای اصلی جهت تحلیل‌های آماری ترافیک شبکه مورد استفاده قرار داد. در ذیل تعاریف پرکاربردترین فیلدهای داده‌ای محاسباتی در تحلیل و پایش ترافیک وب، آورده شده است[۱۳]:

- **HIT(تعداد درخواست‌ها به ازای بازدید یک صفحه)**: هنگامی که کاربر، اقدام به باز کردن یک صفحه می‌کند، مرورگر آن، درخواست HTTP را برای خدمت‌دهنده وب مقصد ارسال می‌کند. وب‌خدمت پاسخ کاربر شامل قالب کلی صفحه و آدرس منابع موردنیاز را ارسال می‌کند. مرورگر کاربر در این مرحله، برای دریافت منابع، درخواست‌های خود را ارسال می‌کند. تمامی منابع موردنیاز در سرایند صفحه دریافتی، به‌طور هم‌زمان و درخواست مابقی منابع، پس‌ازآن ارسال می‌شوند. در اصطلاح به درخواست‌های ارسال شده، برای دریافت منابع موردنیاز برای بارگذاری کامل یک صفحه، HIT گفته می‌شود.
- **Pageviews (تعداد صفحات بازدید شده)**: به تعداد دفعات دیده شدن یک صفحه Page View گفته می‌شود. لذا یک صفحه معادل با یک یا چند HIT می‌باشد اما یک HIT لزوماً معادل یک صفحه نیست. بنابراین این فیلد، تعداد صفحاتی از یک دامنه که توسط کاربران بازدیدکننده مشاهده‌شده است را نگه می‌دارد. (تعداد تکراری بازدیدها نیز به ازای هر دامنه شمارش می‌شود)
- **Unique Pageviews (تعداد صفحات منحصربه‌فرد بازدید شده)**: تعداد صفحات منحصربه‌فردی از یک دامنه(URL) که توسط کاربران بازدیدکننده مشاهده‌شده است. (تعداد تکراری بازدیدها بر اساس آدرس صفحه شمارش نمی‌شود)
- **Unique Visitors (بازدیدکنندگان منحصربه‌فرد)**: تعداد بازدیدکننده‌ها (بر اساس IP) که از یک دامنه و صفحات آن بازدید کرده‌اند. (تعداد تکرار بازدید شمارش نمی‌شود)
- **Visits/Sessions (نشست بازدید)**: تعداد نشست‌های بازدید کاربران از یک دامنه در طول یک روز یا یک بازه زمانی (هفته، ماه و غیره). تفکیک این نشست‌ها از یکدیگر در صورت در اختیار داشتن Session

ID، با محاسبه یکتایی Session ID به ازای هر IP منحصربه‌فرد محاسبه می‌گردد که از دقت بالایی برخوردار است، در غیر این صورت بر اساس یک Interval Time (فاصله زمانی)، این تفکیک صورت می‌گیرد. به‌عنوان نمونه می‌توان تفکیک هر نشست را با نشست قبلی از یک IP منحصربه‌فرد بافاصله زمانی ۳۰ دقیقه‌ای در طول ۱ روز در نظر گرفت.

- **Single Page Visits/Sessions (Singleton - Bounce) – (تعداد نشست‌های شامل بازدید از فقط یک صفحه):** تعداد نشست‌هایی که کاربران بازدیدکننده فقط از یک صفحه از یک دامنه بازدید کرده‌اند.
- **Time on Site (Visit Duration - Session Duration) – (مدت‌زمان نشست بازدید):** میانگین زمان بازدید هر کاربر بازدیدکننده از یک دامنه بر اساس مدت‌زمان نشست‌های بازدید از آن دامنه. در واقع اختلاف زمان بازدید هر یک از تاپل های نشست بازدید (آدرس IP کاربر، شماره نشست، نام دامنه) محاسبه شود و سپس میانگین مجموع آن‌ها به ازای هر دامنه پردازش گردد.

در واقع منظور از ترافیک بازدید کاربران از دامنه‌های وب در شبکه، همان فیلدهای داده‌ای است که از لاگ پروتکل‌های مختلف محاسبه و بدست می‌آیند و به عنوان معیار بازدید کاربران در تحلیل‌های آماری مورد استفاده قرار می‌گیرند. همچنین ممکن است بنا به نیاز و سیاست‌های پایش شبکه، معیارهای ترافیکی دیگری، با ترکیب سایر معیارها ایجاد شوند. چند مورد از این قبیل معیارها در ذیل آورده شده است:

- **Visits (Sessions) per Visitor (تعداد نشست‌ها به ازای هر بازدیدکننده):** میانگین تعداد نشست‌های بازدید هر کاربر بازدیدکننده (IP منحصربه‌فرد) از یک دامنه در طول ۱ روز.
- **Pageviews per Visitor (تعداد صفحات بازدید شده به ازای هر بازدیدکننده):** میانگین تعداد صفحاتی از یک دامنه که توسط کاربران بازدیدکننده مشاهده‌شده است. (تعداد تکراری بازدیدها شمارش می‌شود)
- **Unique Pageviews per Visitor (تعداد صفحات منحصربه‌فرد بازدید شده به ازای هر بازدیدکننده):** میانگین تعداد صفحات منحصربه‌فردی از یک دامنه که توسط کاربران بازدیدکننده مشاهده‌شده است. (تعداد تکراری بازدیدها شمارش نمی‌شود)
- **Pageviews per Visit (Session) (تعداد صفحات بازدید شده به ازای هر نشست بازدید):** میانگین تعداد صفحاتی از یک دامنه که توسط کاربران در یک نشست مشاهده‌شده است. (تعداد تکراری بازدیدها شمارش می‌شود)
- **Unique Pageviews per Visit (Session) (تعداد صفحات منحصربه‌فرد بازدید شده به ازای هر نشست بازدید):** میانگین تعداد صفحات منحصربه‌فردی از یک دامنه که توسط کاربران در یک نشست مشاهده‌شده است. (تعداد تکراری بازدیدها شمارش نمی‌شود)
- **Bounce Rate (نرخ نشست‌های تک‌صفحه‌ای):** نرخ تعداد نشست‌های تک‌صفحه‌ای (Single Page Visits/**Sessions**) نسبت به کل نشست‌های بازدیدکنندگان.
- **Time on Page (Pageviews Duration) – (مدت‌زمان بازدید از هر صفحه):** میانگین زمان بازدید هر کاربر بازدیدکننده از یک صفحه از دامنه.

- **Page Traffic Rank**(رتبه ترافیکی صفحه وب): الگوریتم‌های مختلفی جهت استخراج دقیق‌ترین رتبه‌بندی بر مبنای بازدید کاربران با در نظر گرفتن و وزن دهی معیارهای بازدید روی مجموعه‌ای از صفحات وب می‌تواند تعریف و اعمال شود تا در نهایت صفحه وبی که بیشتر مورد استقبال تعداد بالاتری از کاربران منحصربه فرد و واقعی قرار داشته به عنوان صفحه وب با رتبه ترافیکی اول انتخاب شود.

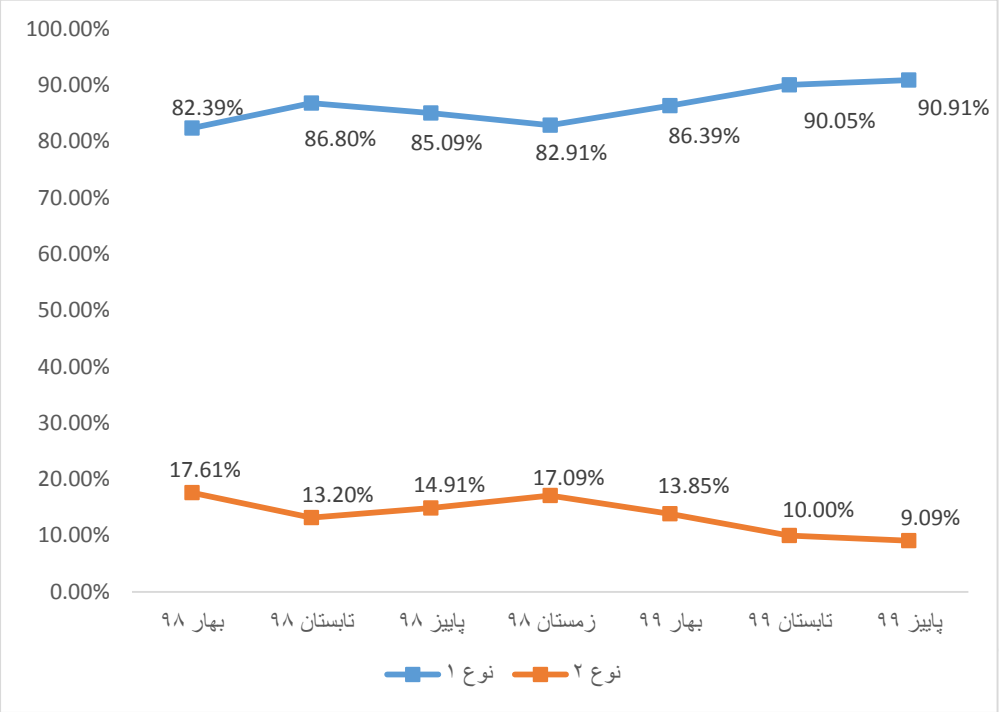
هریک از معیارهای فوق و یا هرگونه معیارهای ترکیبی دیگر مربوط به درخواست بازدید کاربران در شبکه در پروتکل‌های مختلف، می‌توانند به عنوان معیار بازدید در تحلیل‌های آماری در نظر گرفته شوند.

۳-۱-۲-۲ نمونه‌هایی از تحلیل آماری بازدید کاربران شبکه از دامنه‌های وب

با توجه به سیاست‌های پایش شبکه و استخراج داده‌ها و محاسبه معیارها، می‌توان تحلیل‌های آماری مانند موارد زیر را ارائه داد:

- تعداد(مجموع)، میانگین، متوسط بازدید از کل دامنه‌ها یا کاربردها
- تعداد دامنه‌ها یا کاربردهای بازدید شده در بازه زمانی مشخص
- تعداد(مجموع)، میانگین، متوسط، سهم(درصد از کل)بازدید از n کاربرد یا دامنه‌ی برتر در بازه زمانی مشخص
- روند تغییرات تعداد(مجموع)، میانگین، متوسط، سهم(درصد از کل)بازدید از یک تا n کاربرد یا دامنه‌ی برتر در بازه زمانی مشخص
- تعداد(مجموع)، میانگین، متوسط، سهم(درصد از کل)بازدید از کاربردها یا دامنه ها بر اساس معیارهای توصیفی مانند دسته‌بندی موضوعی(یا هر نوع دسته‌بندی دیگر) دامنه‌ها یا کاربردها، موقعیت جغرافیایی دامنه‌ها، موقعیت جغرافیایی کاربران، نوع سیستم عامل یا مرورگر کاربران، پروتکل مورد استفاده در ایجاد ترافیک بازدید(HTTP,HTTPS,DNS,…) و غیره در بازه زمانی مشخص
- تعداد کل دامنه‌های بازدید شده بر اساس هریک از معیارهای توصیفی(مانند مثال‌های مورد قبل) در بازه زمانی مشخص
- لیست رتبه بندی ترافیکی کاربردها یا دامنه‌ها بر اساس تعریف الگوریتم مشخص و در بازه زمانی مشخص

جهت درک بهتر موارد فوق مثال‌های زیر آورده شده است. ***لازم به ذکر است که داده‌های وارد شده در مثال‌ها داده‌های واقعی نیستند و صرفا جهت شفاف سازی موارد فوق، داده‌سازی شده‌اند***. شکل ۲۲ روند سهم بازدید دو دسته موضوعی از دامنه‌ها از کل بازدیدها را در بازه زمانی مشخص نشان می‌دهد.



شکل ۲۲: روند سهم بازدید از دو دسته موضوعی از دامنه‌های وب بازدید شده توسط کاربران از بهار ۹۸ تا پاییز ۹۹

۳-۱-۲-۳ نمونه‌هایی از تحلیل آماری روی حجم ترافیک تبادلی در شبکه

حجم خروجی(آپلود) و حجم ورودی(دانلود) دو شاخص اصلی معیار حجم است که از روی لاگ ترافیک شبکه به ازای دامنه یا کاربرد و یا به ازای کاربران شبکه قابل محاسبه بوده و می‌توان در تحلیل‌های آماری بر اساس معیار حجم از آن‌ها استفاده نمود و با توجه به سیاست‌های پایش شبکه می‌توان تحلیل‌هایی همچون موارد زیر را ارائه داد:

- مجموع حجم ورودی و خروجی کل دامنه‌ها یا کاربردها در لاگ ترافیک شبکه در بازه زمانی مشخص
- مجموع حجم ورودی و خروجی از یک تا n کاربرد یا دامنه در بازه زمانی مشخص
- روند تغییرات حجم ورودی و خروجی از یک تا n کاربرد یا دامنه در بازه زمانی مشخص
- مجموع حجم ورودی و خروجی کاربردها یا دامنه ها بر اساس معیارهای توصیفی مانند دسته‌بندی موضوعی(یا هر نوع دسته‌بندی دیگر) دامنه‌ها یا کاربردها، موقعیت جغرافیایی دامنه‌ها، موقعیت جغرافیایی کاربران، نوع سیستم عامل یا مرورگر کاربران، پروتکل مورد استفاده در ایجاد ترافیک بازدید(HTTP,HTTPS,DNS,...) و غیره در بازه زمانی مشخص
- مجموع حجم ورودی و خروجی به ازای کاربران(یک کاربر یا گروهی از کاربران با ویژگی یکسان مانند رنج IP) در لاگ ترافیک شبکه در بازه زمانی مشخص
- روند تغییرات حجم ورودی و خروجی به ازای کاربران(یک کاربر یا گروهی از کاربران با ویژگی یکسان مانند رنج IP) در لاگ ترافیک شبکه در بازه زمانی مشخص

۳-۱-۲-۴ تحلیل‌های آماری با استفاده از منابع داده‌ای تکمیلی

در برخی موارد جهت ارائه تحلیل‌های آماری دقیق‌تر از داده‌های تکمیلی که از منابع دیگری غیر از لاگ جمع‌آوری شده از شبکه آورده شده‌اند، استفاده می‌شود. این داده‌ها معمولا جهت ارزیابی و مقایسه داده‌های تحلیلی که از لاگ جمع آوری شده است، کاربرد دارد. به عنوان مثال با فرض نمونه گیری یکنواخت از داده‌های تبادلی شرکت ارتباطات زیرساخت در کشور و انجام تحلیل آماری بر روی وب‌سایت‌های بازدید شده و رتبه‌بندی آن‌ها، می‌توان از داده‌های رتبه بندی ارائه شده توسط وب‌سایت جهانی الکسا در زمان و موقعیت جغرافیایی یکسان، مقایسه‌ای را بین داده‌ها انجام داد و نتایج جدید را کشف و ضبط نمود.

همچنین در برخی موارد اطلاعاتی از تحلیل‌های آماری بدست می‌آید که نیاز به ارائه تحلیل‌های توصیفی بیشتری دارد و آن تحلیل‌های توصیفی، داده‌هایی را می‌طلبد که در منبع داده‌ای اصلی وجود ندارد از این رو از فیلدهای داده‌ای تکمیلی از سایر منابع استفاده می‌شود. به عنوان مثال تحلیل‌های آماری نشان می‌دهند که وب‌سایت گوگل بیشترین آمار بازدید در میان کاربران شبکه را داشته است؛ با استفاده از ابزار گوگل ترندز می‌توان کلیدواژه‌ها و موضوعات داغی که مورد جستجوی کاربران بوده است را استخراج نمود و تحلیل توصیفی از آن ارائه داد.

از دیگر نمونه‌های تحلیل‌های مهم در لایه کاربرد تحلیل برنامه‌های کاربردی مورد استفاده کاربران در شبکه است. همانطور که پیش از این ذکر شد با داشتن لاگ پروتکل‌های DPI می‌توان تحلیل‌های آماری و توصیفی مختلفی از بازدید و ترافیک تبادلی کاربران با برنامه‌های کاربردی ایجاد کرد؛ با این حال جهت تدقیق و تکمیل یا ایجاد امکان مقایسه این تحلیل‌ها، می‌توان از اطلاعات فروشگاه‌های آنلاین برنامه‌های کاربردی داخلی و خارجی مانند کافه بازار و پلی استور استفاده نمود. اطلاعات این ابزارها توسط خزشگر و یا API آن‌ها می‌تواند در احتیار قرار بگیرد.

در دسته‌بندی دامنه‌ها همانطور که پیش‌تر اشاره شد، دامنه‌هایی که به هیچ دسته‌ای تعلق ندارند به عنوان دامنه‌های ناشناس برچسب می‌خورند. حال برای درک بهتر این دامنه‌ها می‌توان اطلاعات بیشتری از منابع دیگر در خصوص آن‌ها کسب نمود به عنوان مثال استفاده از سایت WHOIS جهت بررسی وضعیت ثبتی آن دامنه، یکی از این منابع است. به عنوان نتیجه شهودی برخی دامنه‌ها را از ناشناس به ناشناخته می‌توان تفکیک نمود: دامنه‌هایی که غالباً اسامی غیرمعمول آن‌ها، توسط هیچ شخص حقیقی/حقوقی به ثبت نرسیده و یا آزاد شده است و هیچ‌گونه اطلاعات WHOIS و Nameserver در خصوص این دامنه‌ها در پایگاه‌های داده‌های معتبر مربوطه وجود ندارد و طبعاً قابل ترجمه توسط کارگزاران DNS نمی‌باشند؛ بنابراین ماهیت این دامنه‌ها ناشناخته بوده است.

۴ جمع‌بندی

یکی از مهم‌ترین کاربردهای تحلیل ترافیک **شبکه‌های اپراتورهای فراهم کننده خدمات اینترنت و اینترانت به عنوان زیرساخت شبکه ملی اطلاعات**، مدیریت و پایش شبکه‌ها است. پایش و تحلیل ترافیک شبکه ملی اطلاعات را می‌توان از دو منظر بررسی نمود:

۱- پایش ترافیک شبکه به منظور تحلیل و بررسی رفتار شبکه و کمک به مدیریت هوشمند شبکه از کاربردهای تحلیل ترافیک شبکه‌ها است. در واقع با استقرار نرم افزار جمع آوری، ذخیره سازی، پردازش و تحلیل داده مرتبط با همه NMSهای موجود در شبکه و با بهره‌گیری از الگوریتمهای هوش مصنوعی می‌توان پایش و مانیتورینگ جامعی را بر کل شبکه اعمال نمود. در **شبکه ملی اطلاعات** یکی از الزامات تعریف شده، رصد و پایش ترافیک است که به واسطه آن اقداماتی نظیر موارد زیر تعریف شده است:

- توسعه ابزارهای تحلیل جریان ترافیک در سطح اپراتورهای شبکه
- بهینه سازی توزیع ترافیک در شبکه و تمرکز ردیابی (داخلی و بین المللی)
- توسعه سامانه رصد، پایش، تشخیص و تحلیل و مقابله با *برخی* تهدیدات و حوادث شبکه *مانند مدیریت خرابی ها*
- اعمال مدیریت یکپارچه شبکه در مدیریت منابع و تخصیص و بهره برداری بهینه از منابع ملی و نظارت بر آن
- تضمین کیفیت از طریق فراهم آوری امکان مدیریت انتها به انتها در شبکه و ایجاد امکان توسعه رقابت مبتنی بر کیفیت خدمات در بین ارائه دهندگان مختلف
- رصد کمی و کیفی خدمات
- ایجاد قابلیت جمع آوری، پردازش و تحلیل داده‌های عظیم در مورد انواع خدمات و محتوای شبکه و کاربردهای مختلف
- مدیریت ارتباط با مشتریان

۲- پایش ترافیک وب و کاربردها در شبکه اینترنت یا اینترانت در سطوح مختلف جهانی، ملی یا سازمانی به منظور تحلیل رفتار کاربران شبکه و کمک به شناخت بهتر گرایشات کاربران و در نهایت سیاست‌گذاری دقیق‌تر در این راستا است. همانطور که در اهداف رصد و پایش شبکه ملی اطلاعات اقداماتی نظیر ایجاد ابزار تحلیل بازدیدکنندگان و تبلیغات وب برای کسب و کارهای داخلی و خارجی لایه های مختلف **شبکه ملی اطلاعات** و یا ایجاد قابلیت جمع آوری، پردازش و تحلیل داده‌های عظیم در مورد محتوای مورد بازدید و کاربرد کاربران این شبکه، تعریف شده است.

پایش ترافیک شبکه و مدیریت شبکه به واسطه آن، نیازمند تجزیه و تحلیل داده‌های ترافیکی مرتبط است. تجزیه و تحلیل با سؤالاتی همراه است که نشانه سطوح تحلیل و خودکارسازی آن هستند. اولین سؤال این است که چه اتفاقی افتاده، دارد می‌افتد و ممکن است بیافتد؟ این سطح از تجزیه و تحلیل، تحلیل‌های توصیفی نام دارند.

شناسایی داده‌ها و پروتکل‌های انتقال آن در شبکه و روش‌های جمع‌آوری آن‌ها اولین گام برای تجزیه و تحلیل به حساب می‌آید. این داده‌ها اغلب از لایه سوم تا هفتم شبکه‌ها(طبق مدل مرجع OSI) در ترافیک شبکه موجود هستند

و کامل‌ترین اطلاعات پایش را می‌توان از لایه کاربرد که شامل داده‌های کاربران است، استخراج نمود. لذا این گزارش در دو بخش اصلی جمع‌آوری داده‌های ترافیک شبکه و انجام تحلیل‌های توصیفی روی داده‌ها به منظور پایش و مدیریت شبکه ارائه شده است که در ادامه جمع‌بندی از این دو بخش آورده شده است.

۴-۱ جمع‌آوری داده‌های ترافیک شبکه

داده‌های مورد نیاز برای تحلیل از تجهیزات شبکه در لایه‌های مختلف شبکه و در نقاط مختلف آن استخراج می‌شوند. فیلدهای داده‌ای مورد نیاز برای پایش و تحلیل شبکه با نیازها و سیاست‌های تحلیل ارتباط مستقیم دارد. فیلدهای داده‌ای از بسته‌ها و پیام‌ها در لایه‌های مختلف شبکه استخراج می‌شوند که آن‌ها نیز در قالب پروتکل‌های مشخص انتقال می‌یابند. پروتکل‌های انتقال فیلدهای داده‌ای مشخصی دارند که در هنگام ارتباط و انتقال پیام مقداردهی می‌شوند؛ از این فیلدها برای پایش و تحلیل ترافیک شبکه‌ها می‌توان استفاده نمود. فیلدهای داده‌ای لایه کاربرد بیشترین و کاربردی‌ترین اطلاعات تحلیل را در اختیار کاربران قرار می‌دهد.

جمع‌آوری داده‌ها اغلب با استخراج لاگ فایل‌ها صورت می‌پذیرد. اطلاعاتی که لاگ در اختیار می‌گذارد، یک منبع بسیار مهم برای کسب دانش تحلیل است. به عنوان مثال با استفاده از لاگ یک سرور وب، به راحتی می‌توان تشخیص داد که چه تعداد کاربر و از چه موقعیت‌های جغرافیایی به فایل خاصی دسترسی داشته‌اند و یا با تجمیع لاگ‌های مربوط به ترافیک تبادلی تجهیزات شبکه می‌توان تشخیص داد که چطور می‌توان توزیع بار ترافیکی یا مدیریت منابع را بهینه ساخت و یا حتی پیش‌بینی کرد که در آینده ممکن است چه خطراتی برای شبکه وجود داشته‌باشد. نکته‌ای که بسیار حائز اهمیت می‌باشد این است که جمع‌آوری داده‌ها باید به‌گونه‌ای انجام گیرد که با کل داده‌های شبکه را شامل شود یا توزیع یکنواختی از داده‌ها که به عنوان مثال از نمونه های تصادفی از کل مبادی داده در تمام ساعات شبانه روز تشکیل شده باشد. در مواردی که حجم داده‌ها بسیار بالا است و امکان جمع‌آوری کل داده‌های شبکه وجود ندارد(مانند داده‌های شبکه‌ی زیرساختی کشور) می‌توان از روش‌های نمونه‌گیری استفاده کرد. در این گزارش معرفی روش‌های جمع‌آوری داده و فیلدهای داده‌ای قابل استخراج از هر روش از دو منظر جمع‌آوری داده‌های مدیریت شبکه و داده‌های پایش دامنه‌های وب مورد بازدید کاربران بررسی شده است.

۴-۱-۱جمع‌آوری داده‌های مدیریت شبکه

پروتکل‌های مدیریت شبکه به همراه ابزارها و سیستم‌های مدیریت شبکه، داده‌ها و اطلاعات زیادی از وضعیت شبکه را در اختیار تحلیلگران داده‌های ترافیک شبکه قرار می‌دهند. در ادامه به جمع‌بندی روش‌های جمع‌آوری داده برای مدیریت شبکه و فیلدهای داده‌ای قابل استخراج از هر روش پرداخته شده است.

۴-۱-۱-۱ جمع‌آوری لاگ از نودهای سوئیچ و روتر شبکه

بنابراین استخراج لاگ این گره‌های مهم شبکه که از لایه ۳ تا ۷ شبکه می‌توانند داده‌ها را انتقال دهند(اغلب در لایه ۳ فعالیت دارند)، جهت تحلیل و پایش ترافیک شبکه، داده‌های مفید و موثری را در اختیار می‌گذارند. نحوه استخراج لاگ از این تجهیزات می‌تواند برای برخی از این تجهیزات که قابلیت لاگ گیری دارند، استفاده از لاگ آن‌ها باشد و برای برخی دیگر با قرار دادن یک سرور بک آپ و ارسال یک نسخه از کلیه ترافیک عبوری به آن صورت پذیرد.

۴-۱-۱-۲ جمع‌آوری لاگ از سیستم‌های مدیریت شبکه

پروتکل‌های مدیریت شبکه مانند پروتکل NetFlow به همراه ابزارها و سیستم‌های مدیریت شبکه، داده‌ها و اطلاعات زیادی از وضعیت شبکه را در اختیار تحلیلگران داده‌های ترافیک شبکه قرار می‌دهند.

۴-۱-۲جمع‌آوری داده‌های ترافیک وب

فیلدهای داده‌ای پروتکل‌های مطرح ترافیک وب در لایه کاربرد شامل HTTP،HTTPS،DNS،DPI در این گزارش معرفی شده‌اند. بر اساس مدل OSI هر بسته ارسالی در شبکه از ترکیب payload+header تشکیل می‌شود و یک واحد انتقال بسته در شبکه را می‌سازد. از مهم‌ترین فیلدهای موجود در این سرآیندها که در پایش ترافیک وب موثر هستند، می‌توان به موارد ذیل اشاره نمود:

- IP مبدأ و IP مقصد (سرآیند IP)
- PORT مبدأ و PORT مقصد، زمان و تاریخ درخواست (سرآیند TCP)
- نام دامنه، آدرس URL، شناسه عامل کاربری (User-agent)، نوع محتوا (Content-type) – (سرآیند پروتکل HTTP)

در این گزارش نمونه‌هایی از روش‌ها و نقاط استخراج داده‌های ترافیک وب در شبکه‌های فراهم کنندگان خدمات اینترنت و اینترانت و شبکه‌های سازمانی بزرگ، شرح داده شدند که در ادامه مختصری از هریک آورده شده است:

۴-۱-۲-۱ جمع‌آوری لاگ سرورهای میزبانی مراکز داده

یکی دیگر از نقاطی که لاگ ترافیک عبوری آن در پایش ترافیک شبکه‌ها، بالاخص شبکه اینترنت و اینترانت مفید و موثر است، لاگ سرورهای مراکز داده است. ترافیک کاربران شبکه های بزرگ جهت بازدید از وب به سوی اپراتورهای مراکز داده که میزبان سرورهای وب هستند، روانه می‌شوند؛ با بررسی ترافیک بازدید کاربران از دامنه‌های مختلف می‌توان اطلاعات ارزشمندی از گرایش کاربران در بازدید از آن‌ها و همچنین اطلاعات وضعیت شبکه و ترافیک آن بدست آورد.

۴-۱-۲-۲ جمع‌آوری لاگ ربات‌ها و اسکریپت‌ها

ربات‌ها، افزونه‌های مرورگرها و اسکریپت‌های وب یکی دیگر از روش‌های جمع‌آوری داده‌های ترافیکی بازدید از وب‌سایت‌ها است که در صورت حفظ جامعیت و تنوع می‌تواند در تحلیل ترافیک وب مورد استفاده قرار بگیرد. اسکریپت‌های وب توسط ابزارهای تحلیل وب مانند گوگل آنالیتیکز توسط صاحبان وبسایت‌ها روی صفحات وب ایشان قرار می‌گیرد و اطلاعات تحلیلی در اختیار آن‌ها می‌گذارد. همچنین ارگان‌های مرجعی در سراسر دنیا وجود دارد که اسکریپتی جهت ارزیابی فعالیت‌های وب به عنوان گواهینامه‌های مختلف به خواست صاحبان وب در وبسایت‌ها قرار می‌دهند که ترافیک آن وبسایت‌های دارای اسکریپت اطلاعات دقیق و سطح بالایی از بازدید کاربران از آن وب‌سایت‌ها را در اختیار تحلیل‌گران وب قرار می‌دهد. به عنوان مثال در ایران وزارت صمت اسکریپت نماد اعتماد الکرونیکی را ارائه می‌دهد و وزارت فرهنگ و ارشاد اسلامی اسکریپت نماد رسانه‌های دیجیتال را که یک ارزیابی محتوایی است، ارائه می‌دهد. در ادامه به شناخت بیشتری از لاگ‌های اسکریپتی پرداخته شده است.

۴-۱-۲-۳ جمع‌آوری لاگ از فایروال‌ها و سیستم‌های امنیت شبکه

در اکثر طراحی‌ها تمام ترافیک ورودی و خروجی یک شبکه از فایروال‌های آن می‌گذرد و فایروال‌ها در شبکه‌ها(مانند شبکه اپراتورهای فراهم کننده خدمات اینترنت، اینترانت، داده و غیره) یکی از مناسب‌ترین نقاط برای واکشی داده‌های ترافیکی شبکه در لایه‌های مختلف هستند و می‌توان توسط آن‌ها ترافیک ورودی و خروجی به شبکه را پایش نمود. از دیگر خدمت‌های مرتبط با فایروال‌ها و مباحث امنیت شبکه، استخراج لاگ کاربردهای DPI است. پروکسی‌ها که غالباً به‌عنوان فیلترشکن مورد استفاده قرار می‌گیرند. برنامه‌های کاربردی مانند برخی پیام‌رسان‌ها یا شبکه‌های اجتماعی که کاربران زیادی داشته و ترافیک بالایی را به خود اختصاص داده‌اند و یا به دلایل دیگری حائز اهمیت هستند، پروتکل‌های شبکه مانند HTTPS و DNS و FTPS و غیره که احتمالاً به‌طور ترکیبی با قوانین دیگری جهت نیل به اهداف مشخص و همچنین برخی دامنه‌هایی که به دلایل خاصی حائز اهمیت هستند، در DPI برخی شبکه‌های فراهم کننده خدمات اینترنت در کشور، تعریف شده‌اند.

۴-۱-۲-۴ جمع‌آوری لاگ از سرورهای DNS

در یک ارتباط انتها به انتها در شبکه اینترنت و اینترانت از تجهیز کاربر گرفته تا مراکز داده و اپراتورهای فراهم کننده خدمات ارتباطی از پروتکل DNS جهت تبدیل نام دامنه به آدرس IP و مسیریابی پیام‌ها استفاده می‌کنند. بنابراین سرورهای DNS شامل اطلاعات مفیدی برای تحلیل ترافیک این شبکه‌ها هستند.

۴-۲ پایش و مدیریت شبکه با ارائه تحلیل‌های توصیفی بر روی لاگ ترافیک شبکه

تحلیل‌های آماری و مبتنی بر هوش مصنوعی متنوعی می‌تواند به پایش ترافیک شبکه در راستای مدیریت بهتر شبکه و سیاست‌گذاری برای آن، کمک کند. همان‌طور که در ابتدای بخش جمع‌بندی ذکر شد، به تحلیل‌ها نیز می‌توان همانند داده‌ها از دو منظر پایش و مدیریت شبکه و پایش ترافیک وب نگاه کرد.

۴-۲-۱ تحلیل‌های توصیفی به منظور پایش و مدیریت شبکه

عناوین این قبیل تحلیل‌ها جهت آشنایی با مفهوم و ذکر نمونه در گزارش ذکر شدند:

- تحلیل دسته‌بندی ترافیک بر اساس معیارهای مشخص مانند دسته‌بندی موضوعی کاربردهای شبکه
- تحلیل احساسات روی داده‌های کیفیت خدمت و میزان رضایتمندی کاربران شبکه
- تحلیل‌های مبتنی بر گراف به منظور شناسایی الگوی ترافیک و کمک به شناخت وضعیت و باگ‌های شبکه
- پایش وضع موجود و ارائه تحلیل‌های پیش‌بینی وضعیت آینده شبکه به منظور ایجاد سیستم‌های یکپارچه مدیریت شبکه، برای جلوگیری از وقوع مشکلات، خرابی‌ها، مدیریت شرایط بحرانی و ازدحام ترافیک (Congestion)
- تحلیل میزان بهره‌وری کل شبکه با استفاده از داده‌های آموزشی
- تحلیل روندهای تغییر الگوهای ترافیک مشتریان و وفق پذیری با آن
- تحلیل های پایداری شبکه و نظایر آن

شرح راه‌کارها و ارائه الگوریتم‌های هوش مصنوعی، یا تعریف روش‌ها و نمونه‌های آماری به ازای تمام موارد تحلیلی نامبرده فوق و سایر موارد مشابه به دلیل حجم بالای کار و همچنین نیاز به شناخت شبکه، سیاست‌های سازمانی و

داده‌های گردآوری شده از آن شبکه در این گزارش نمی‌گنجد و بر اساس نیازهای موجود در ارگان‌ها می‌تواند در قالب یک پروژه تعریف شود.

۴-۲-۲تحلیل‌های توصیفی ترافیک وب

تحلیل توصیفی ترافیک ایجاد شده توسط کاربران در بازدید از دامنه‌های وب با دو روش اصلی تحلیل‌های مبتنی بر هوش مصنوعی و تحلیل‌های آماری می‌توان انجام داد.

۴-۲-۲-۱ تحلیل‌های مبتنی بر هوش مصنوعی

نمونه‌هایی از این تحلیل‌ها برای پایش ترافیک وب شامل موارد زیر است:

- تحلیل احساسات کاربران در نظرات ارائه شده
- تحلیل تعاملات کاربران، وب‌سایت‌ها و محتواها
- انواع تحلیل‌های شبکه‌های اجتماعی
- تحلیل دسته‌بندی: منظور از دسته‌بندی در این گزارش، دسته بندی محتوای متنی تحت وب است که به عنوان نمونه در بخش تحلیل‌های توصیفی ترافیک وب، روش کلی آن شرح داده شده است. بدین صورت که ابتدا باید موضوع مربوط به هرکدام از این محتواهای متنی تشخیص داده شده و سپس بر اساس این موضوعات دسته‌بندی آن‌ها انجام شود. انجام تحلیل‌های مختلف مانند تحلیل دسته‌بندی دامنه‌های وب نیاز به داشتن محتوای هریک از نام دامنه‌های وب لاگ شده در ترافیک شبکه است. یکی از رویکردهای اصلی در حوزه گردآوری داده از وب، خزش وب می‌باشد.برای دسته‌بندی دامنه‌ها مبتنی بر محتوای متنی، به‌طورکلی می‌توان آن‌ها را به دو دسته تقسیم نمود. دامنه‌هایی که دارای محتوای متنی هستند و دامنه‌هایی که بدون محتوای متنی هستند. ازاین‌رو در صورت وجود محتوای متنی قابل‌اتکا می‌توان توسط الگوریتم‌های یادگیری ماشین با استفاده از روش های با نظارت و از روش‌های مبتنی بر قوانین برای دامنه‌های فاقد محتوا، دسته‌بندی‌ها را انجام داد.

۴-۲-۲-۲ تحلیل‌های آماری

هدف از این تحلیل‌ها، کشف و نمایش الگوهای ترافیکی بر گرفته از تحلیل آماری رفتار شبکه و کاربران است. تحلیل‌های آماری روی داده های خام و یا اطلاعات حاصل از سایر تحلیل‌های توصیفی و یا به صورت ترکیبی از آن‌ها انجام می‌شود. به عنوان مثال تحلیل آماری روی اطلاعات دسته‌بندی دامنه‌های وب و همچنین داده‌های بازدید از آن‌ها می‌تواند جهت‌گیری و علایق افراد در بازدید از دامنه‌های وب در آن شبکه را نشان دهد. با دیدگاه مدیریت شبکه، تحلیل آماری ترافیک شبکه در لایه کاربرد ابزاری قدرتمند جهت استخراج الگوهای رفتاری کاربران در شبکه است و امکان رصد عمیق‌تر و دقیق‌تر جریانات آن شبکه را در اختیار مسئولان و سیاست‌گذاران آن شبکه می‌گذارد.

- ***تحلیل آماری روی بازدید کاربران***

منظور از بازدید کاربران در لاگ ترافیک شبکه فیلدهای داده‌ای محاسباتی یا معیارهایی هستند که از روی فیلدهای داده‌ای پایه (در لاگ شبکه) بدست می‌آیند. به عنوان مثال فیلد محاسباتی Unique Pageviews

(تعداد صفحات منحصربه‌فرد بازدید شده) که تعداد صفحات منحصربه‌فردی از یک دامنه(URL) که توسط کاربران بازدیدکننده مشاهده‌شده است (تعداد تکراری بازدیدها بر اساس آدرس صفحه شمارش نمی‌شود).

حال با توجه به سیاست‌های پایش شبکه و استخراج فیلدهای داده‌ای و محاسبه معیارها می‌توان تحلیل‌های آماری مانند موارد زیر را ارائه داد:

- تعداد(مجموع)، میانگین، متوسط بازدید از کل دامنه‌ها یا کاربردها
- تعداد دامنه‌ها یا کاربردهای بازدید شده در بازه زمانی مشخص
- تعداد(مجموع)، میانگین، متوسط، سهم(درصد از کل)بازدید از n کاربرد یا دامنه‌ی برتر در بازه زمانی مشخص
- روند تغییرات تعداد(مجموع)، میانگین، متوسط، سهم(درصد از کل)بازدید از یک تا n کاربرد یا دامنه‌ی برتر در بازه زمانی مشخص
- تعداد(مجموع)، میانگین، متوسط، سهم(درصد از کل)بازدید از کاربردها یا دامنه ها بر اساس معیارهای توصیفی مانند دسته‌بندی موضوعی(یا هر نوع دسته‌بندی دیگر) دامنه‌ها یا کاربردها، موقعیت جغرافیایی دامنه‌ها، موقعیت جغرافیایی کاربران، نوع سیستم عامل یا مرورگر کاربران، پروتکل مورد استفاده در ایجاد ترافیک بازدید(HTTP,HTTPS,DNS,…) و غیره در بازه زمانی مشخص
- تعداد کل دامنه‌های بازدید شده بر اساس هریک از معیارهای توصیفی(مانند مثال‌های مورد قبل) در بازه زمانی مشخص
- لیست رتبه بندی ترافیکی کاربردها یا دامنه‌ها بر اساس تعریف الگوریتم مشخص و در بازه زمانی مشخص

- ***تحلیل آماری روی حجم ترافیک تبادلی در شبکه***

حـجم خروجی(آپلود) و حجم ورودی(دانلود) دو شاخص اصلی معیار حجم است که از روی لاگ ترافیک شبکه به ازای دامنه یا کاربرد و یا به ازای کاربران شبکه قابل محاسبه بوده و می‌توان در تحلیل‌های آماری بر اساس معیار حجم از آن‌ها استفاده نمود و با توجه به سیاست‌های پایش شبکه می‌توان تحلیل‌هایی همچون موارد زیر را ارائه داد:

- مجموع حجم ورودی و خروجی کل دامنه‌ها یا کاربردها در لاگ ترافیک شبکه در بازه زمانی مشخص
- مجموع حجم ورودی و خروجی از یک تا n کاربرد یا دامنه در بازه زمانی مشخص
- روند تغییرات حجم ورودی و خروجی از یک تا n کاربرد یا دامنه در بازه زمانی مشخص
- مجموع حجم ورودی و خروجی کاربردها یا دامنه ها بر اساس معیارهای توصیفی مانند دسته‌بندی موضوعی(یا هر نوع دسته‌بندی دیگر) دامنه‌ها یا کاربردها، موقعیت جغرافیایی دامنه‌ها، موقعیت جغرافیایی کاربران، نوع سیستم عامل یا مرورگر کاربران، پروتکل مورد استفاده در ایجاد ترافیک بازدید(HTTP,HTTPS,DNS,…) و غیره در بازه زمانی مشخص
- مجموع حجم ورودی و خروجی به ازای کاربران(یک کاربر یا گروهی از کاربران با ویژگی یکسان مانند رنج IP) در لاگ ترافیک شبکه در بازه زمانی مشخص

- روند تغییرات حجم ورودی و خروجی به ازای کاربران(یک کاربر یا گروهی از کاربران با ویژگی یکسان مانند رنج IP) در لاگ ترافیک شبکه در بازه زمانی مشخص

- ***تحلیل‌های توصیفی با استفاده از منابع داده‌ای تکمیلی***

در برخی موارد جهت ارائه تحلیل‌های آماری دقیق‌تر از داده‌های تکمیلی که از منابع دیگری غیر از لاگ جمع‌آوری شده از شبکه آورده شده‌اند، استفاده می‌شود. این داده‌ها معمولا جهت ارزیابی و مقایسه داده‌های تحلیلی که از لاگ جمع آوری شده است، کاربرد دارد. به عنوان مثال با فرض نمونه گیری یکنواخت از داده‌های تبادلی شرکت ارتباطات زیرساخت در کشور و انجام تحلیل آماری بر روی وب‌سایت‌های بازدید شده و رتبه‌بندی آن‌ها، می‌توان از داده‌های رتبه بندی ارائه شده توسط وب‌سایت جهانی الکسا در زمان و موقعیت جغرافیایی یکسان، مقایسه‌ای را بین داده‌ها انجام داد و نتایج جدید را کشف و ضبط نمود.

همچنین در برخی موارد اطلاعاتی از تحلیل‌های آماری بدست می‌آید که نیاز به ارائه تحلیل‌های توصیفی بیشتری دارد و آن تحلیل‌های توصیفی، داده‌هایی را می‌طلبد که در منبع داده‌ای اصلی وجود ندارد از این رو از فیلدهای داده‌ای تکمیلی از سایر منابع استفاده می‌شود. به عنوان مثال می‌توان از ابزار گوگل ترندز در کنار تحلیل روند بازدید یا حجم تبادلی موتورجستجوی گوگل و یا به‌کارگیری اطلاعات برنامه‌های کاربردی محبوب کاربران در فروشگاه‌های آنلاین برنامه‌های کاربردی مانند پلی استور در کنار تحلیل بازدید از برنامه‌های کاربردی پروتکل DPI و ازاین قبیل موارد، به عنوان منابع داده‌ای تکمیلی نام برد.

مراجع

۱ HTTPS://en.wikipedia.org/wiki/OSI_model

۲ HTTPS://en.wikipedia.org/wiki/Syslog

۳ HTTPS://en.wikipedia.org/wiki/Simple_Network_Management_Protocol

۴ “Book: Uri's Netflow Traffic Logs' Behavioral Analysis and Monitoring Visualization Tool”,Gebregiorgis, Semhar Kessete.University of Rhode Island. ProQuest Dissertations Publishing, ۲۰۱۸. ۱۰۷۹۳۷۳۶.

۵ Hypertext Transfer Protocol -- HTTP/۱,۱. (n.d.). Retrieved October ۳۱, ۲۰۱۷, from HTTPS://tools.ietf.org/html/rfc۲۶۱۶section-۱۴,۲۳

۶ R. ۶۱۰۱, "The SSL Protocol Version ۳,۰," August ۲۰۱۱. [Online]. Available: HTTPS://tools.ietf.org/html/rfc۶۱۰۱.

۷ DPI Solutions in Practice: Benchmark and Comparison,Tommaso Rescio Politecnico di Torino, Italy; Thomas Favale; Francesca Soro; Marco Mellia; Idilio Drago,۲۰۲۱ IEEE Security and Privacy Workshops (SPW)

۸ E. KARTMANN, " Simple DNS Resolver v۱,۵," ۲۰۰۲. [Online]. Available: HTTP://www.kartmann.org/software/SimpleDNSClient/Readme.htm.

۹ Deep Learning and Data Sampling with Imbalanced Big Data,Justin M. Johnson; Taghi M. Khoshgoftaar,۲۰۱۹ IEEE ۲۰th International Conference on Information Reuse and Integration for Data Science (IRI)

۱۰ Measuring Techniques and Data Analysis,Michael C. H. McKubre,Digby D. Macdonald,Brian Sayers,J. Ross Macdonald, ۲۵ March ۲۰۱۸

۱۱ Szegedy, Christian, et al. "Inception-v۴, inception-resnet and the impact of residual connections on learning." AAAI. Vol. ۴. ۲۰۱۷.

۱۲ Sabbah, Thabit, et al. "Modified frequency-based term weighting schemes for text classification." Applied Soft Computing ۵۸ (۲۰۱۷): ۱۹۳-۲۰۶.

۱۳ Web Analytics Definitions. (n.d.). Retrieved October ۳۱, ۲۰۱۷, from HTTP://www.logaholic.com/manual/references/web-analytics-definitions/#PageViews

نشانی: تهران، انتهای کارگر شمالی، پژوهشگاه ارتباطات و فناوری اطلاعات، معاونت پژوهش و توسعه ارتباطات علمی

تلفن: ۰۲۱-۸۸۶۳۰۳۵۵

نمابر: ۰۲۱-۸۸۶۳۰۳۵۶