مروری بر حمله‌های انگشت‌نگاری تارنما

مریم طائبی[1]، علی بهلولی*[2] و مرجان کائدی[3]

[1] کارشناسی ارشد مهندسی کامپیوتر، دانشگاه اصفهان
m.taebi@eng.ui.ac.ir
[2] استادیار دانشکدهٔ مهندسی کامپیوتر، دانشگاه اصفهان
bohlooli@eng.ui.ac.ir
[3] استادیار دانشکدهٔ مهندسی کامپیوتر، دانشگاه اصفهان
kaedi@eng.ui.ac.ir

چکیده

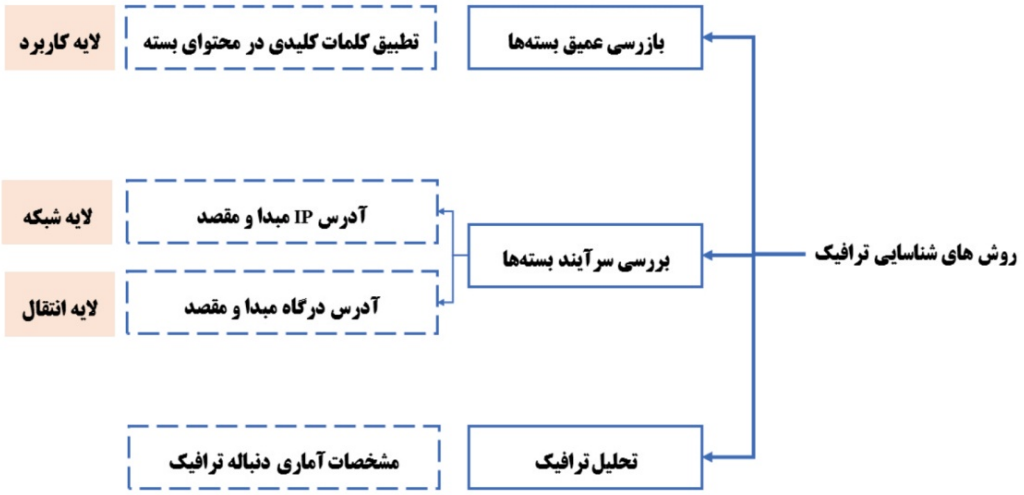
در حملات انگشت‌نگاری تارنما، مقصد ارتباط کاربر بدون رمزگشایی محتوای ترافیک، با استفاده از روش‌های تحلیل ترافیک شناسایی می‌شود. در این حملات، به‌طورمعمول کاربران از یکی از فنـــاوری‌های روز (شبکه‌های گم‌نامی، پراکسی‌ها یا VPNها) برای پنهان‌کـــردن محتوای ترافیک و مقصد واقعی خود استفاده می‌کنند. با استخراج مجموعه‌ای از ویژگی‌ها از دنبالهٔ ترافیک ورودی، حمله آغاز می‌شود؛ سپس داده‌ها پیش‌پردازش می‌شوند و درنهایت، از یک الگوریتم یادگیری ماشین برای شناسایی مقصد ترافیک کاربر استفاده می‌شود. پژوهش‌های متنوعی در زمینهٔ شناسایی مقصد ترافیک یا به‌طور واضح‌تر، تعیین صفحهٔ وب مرورشده توسط کاربر با بهره‌گیری از روش‌های شناخته‌شدهٔ دسته‌بندی در حوزه یادگیری ماشین انجام گرفته‌است. در این مقاله، مروری جامع بر روش‌های انگشت‌نگاری تارنما انجام می‌شود که در آن، پژوهش‌های بالا بر اساس ویژگی‌های مورد استفاده برای انگشت‌نگاری، دسته‌بندی می‌شوند. این نوع دسته‌بندی با دیدگاهی تازه انجام می‌شود که بنابر دانش ما، تاکنون بر روی پژوهش‌های یادشده صورت نگرفته است.

واژگان کلیدی: حملهٔ انگشت‌نگاری تارنما، تحلیل ترافیک، یادگیری ماشین، دسته‌بندی، استخراج ویژگی

۱- مقدمه

مرور صفحات وب توسط یک کاربر، به تولید داده‌های متنوعی می‌انجامد. موجودیت‌های مختلفی در شبکه قادرند به جمع‌آوری این داده‌ها و استخراج اطلاعات از آن‌ها بپردازند. این کار با اهداف گوناگونی مانند تحلیل رفتار کاربران برای مسائل اجتماعی و تعیین الگوی علاقه‌مندی کاربران برای تبلیغات انجام می‌شود [۱-۲]. یکی از اطلاعات موردعلاقه، تعیین مقصد ترافیک کاربر در حین مرور وب یا به عبارت دیگر، شناسایی صفحه وب مرورشده توسط کاربر است. روش‌های مختلفی برای دست‌یابی به این هدف ارائه شده‌اند (شکل ۱). یکی از این روش‌ها، بازرسی عمیق بسته برای کسب اطلاعات از محتوای بسته‌ها (در لایهٔ کاربرد) و در نتیجه مقصد نهایی ارتباط کاربر است [۳]. در این روش‌ها، بسته‌های جریان ترافیک، برای یافتن الگوها یا کلمات کلیدی خاص بررسی می‌شوند [۴]. این الگوها می‌توانند شامل URL یک صفحه وب یا کلمات کلیدی خاص موجود در URL باشند [۵]. با این حال، استفاده از ترافیک رمزشده در طول یک ارتباط (به‌عنوان مثال استفاده از پروتکل HTTPS)، مانع از دسترسی به محتوای بسته‌ها و در نتیجه مانع استفاده از روش‌های بازرسی عمیق بسته برای شناسایی مقصد ارتباط می‌شود. روش دیگر برای کشف ارتباط بین فرستنده و گیرنده، استفاده از چهارتایی نشانی IP مبدأ، شمارهٔ درگاه مبدأ، نشانی IP مقصد، شمارهٔ درگاه مقصد است. معرفی پراکسی‌ها، روش‌های تونل‌زنی ترافیک و شبکه‌های mix[1] (که با هدف پنهان کردن مبدأ، مقصد و محتوای ارتباط کاربران مختلف به کار می‌رود) و استفادهٔ روزافزون کاربران از آن، بهره‌گیری از نشانی‌های موجود را نیز در لایهٔ شبکه و لایهٔ انتقال برای تعیین مقصد ارتباط با محدودیت مواجه کرده‌است [۵]. در این شرایط، تشخیص صفحهٔ وب با استفاده از روش‌های تحلیل ترافیک و به‌کارگیری یکی از روش‌های دسته‌بندی رایج در حوزهٔ یادگیری ماشین انجام می‌شود.

[1] شبکه‌های mix نمونه‌ای از شبکه‌های گم‌نامی هستند که در آن‌ها برقراری ارتباط بین دو نقطه از طریق زنجیره‌ای از مسیریاب‌ها در یک شبکهٔ روی‌هم‌گذاری، بین مبدأ و مقصد ارتباط انجام می‌شود.



(شکل-۱): روش‌های شناسایی مقصد یک دنبالهٔ ترافیک

تحلیل ترافیک به معنای استنتاج اطلاعات از مشخصات آماری یک دنبالهٔ ترافیک و حملهٔ انگشت‌نگاری تارنما یک حالت خاص از تحلیل ترافیک با هدف شناسایی مقصد ارتباط کاربر است.

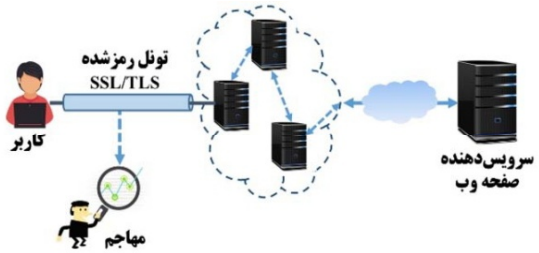
حملهٔ انگشت‌نگاری تارنما را یک مهاجم محلی و منفعل انجام می‌دهد [۶]. منظور از مهاجم محلی و منفعل، گره‌ای است که در مسیر کاربر و پیش از پراکسی یا نخستین گره، در یک شبکهٔ mix (مطابق شکل ۲) قرار می‌گیرد، فقط به جمع‌آوری داده‌های ترافیک می‌پردازد و دخل و تصرفی در دنبالهٔ ترافیک ایجاد نمی‌کند. این مسأله باعث می‌شود که مهاجم برای گره‌های موجود در مسیر قابل شناسایی نباشد.

برای تشخیص مقصد ترافیک کاربر، دو رویکرد قابل تصور است: در یک رویکرد، کاربر تنها محدود به ملاقات مجموعه‌ای از صفحات وب تحت‌عنوان مجموعهٔ «تحت نظارت[1]» مهاجم است. یک مهاجم، در صورتی قدرت‌مند تلقی می‌شود که با دریافت نمونهٔ ورودی بتواند با دقت مناسبی تشخیص دهد کاربر کدام‌یک از صفحات موجود در مجموعهٔ تحت‌نظارت را بازدید کرده‌است. این تنظیمات، تنظیمات جهان-بسته[2] نامیده‌می‌شود [۷]. در ارزیابی این رویکرد، اساساً توانایی دسته‌بندی الگوریتم‌ها مدّنظر است. در رویکرد دیگر (رویکرد جهان-باز[3])، کاربر اجازه ملاقات کلیهٔ صفحات وب را دارد [۸]. دنباله‌های ترافیکی که در مجموعهٔ تحت‌نظارت مهاجم قرار نداشته‌باشند، مجموعهٔ «بدون نظارت[4]» یا ترافیک پس‌زمینه را تشکیل می‌دهند [۹،۱۰]. کلیهٔ صفحات موجود در مجموعهٔ تحت‌نظارت را می‌توان در زبان یادگیری ماشین، طبقهٔ مثبت و سایر صفحات را طبقهٔ منفی نامید. در این تنظیمات، مهاجم باید با آموزش دسته‌بند برروی مجموعهٔ صفحات تحت‌نظارت، تشخیص دهد آیا کاربر یکی از صفحات موجود را در مجموعهٔ تحت‌نظارت (کلاس مثبت) ملاقات کرده است یا صفحهٔ مرورشده در مجموعهٔ بدون‌نظارت (طبقهٔ منفی) قرار دارد [۸]. در یک حملهٔ دقیق‌تر، مهاجم قادر است نوع صفحهٔ ملاقات‌شده توسط کاربر را نیز در مجموعهٔ تحت‌نظارت تشخیص دهد [۹].

در این مقاله، مروری جامع بر روش‌های انگشت‌نگاری تارنما انجام می‌شود. در این مرور، پژوهش‌های بالا برای نخستین بار بر اساس ویژگی‌های مورداستفاده برای انگشت‌نگاری، دسته‌بندی می‌شوند؛ بنابراین مطالعهٔ این پژوهش، به آشنایی با سیر تحول حملات انگشت‌نگاری تارنما در طول زمان، چالش‌های آن‌ها و شناخت مسیر آینده انگشت‌نگاری تارنما کمک می‌کند.

در ادامهٔ این مقاله، در بخش ۲ فرایند مرور وب به‌اختصار توضیح داده‌می‌شود. در بخش ۳ حملات انگشت‌نگاری تارنمای معرفی‌شده در پژوهش‌های پیشین در هفت زیربخش دسته‌بندی و بررسی می‌شود. همان‌طور که یاد شد، دسته‌بندی حملات بر اساس ویژگی‌های مورداستفاده برای شناسایی تارنما انجام می‌شود. درنهایت در بخش ۴ مقاله جمع‌بندی می‌شود.

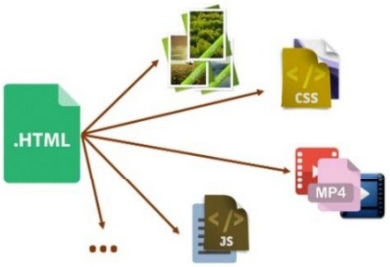
[1] Monitored set
[2] Closed-world
[3] Open-World
[4] Non-Monitored set



(شکل-۲): سناریوی حملهٔ انگشت‌نگاری تارنما

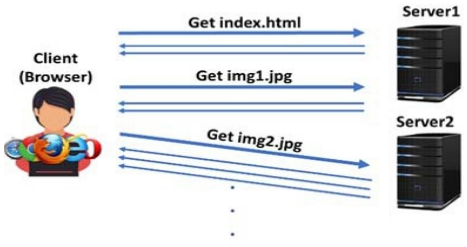
۲- مرور وب

یک صفحهٔ وب، متشکل از یک سند HTML است که ساختار کلی صفحه را تعیین می‌کند و می‌تواند حاوی تعدادی شیء[1] باشد (شکل ۳-الف). هر شیء می‌تواند یکی از انواع فایل مانند یک تصویر JPEG یا PNG، یک فایل ویدئویی، اسکریپت جاوا، کد CSS یا انواع دیگر فایل باشد [۱۱]. زمانی‌که کاربر قصد مشاهدهٔ یک صفحه وب را داشته‌باشد، نشانی URL آن صفحه را در مرورگر خود وارد می‌کند. مرورگر یک اتصال TCP با سرویس‌دهندهٔ مربوط برقرار کرده، سپس با ارسال پیام GET به سرویس‌دهندهٔ موردنظر، نخستین شیء صفحه، یعنی سند HTML را دریافت می‌کند. مرورگر پس از پردازش صفحه، فهرستی از اشیای صفحه را استخراج می‌کند. در پروتکل HTTP، هر شیء معادل یک منبع[2] با شناسه یکتای URI است. برای دریافت هر شیء، نخست یک اتصال به سرویس‌دهندهٔ دارندهٔ آن منبع برقرار می‌شود؛ سپس درخواست GET حاوی شناسهٔ منبع به سرویس‌دهنده ارسال می‌شود، به عبارت دیگر، ممکن است منابع مختلف یک صفحه در سرویس‌دهنده‌های مختلفی قرار گرفته‌باشند و برای دریافت هرکدام لازم است یک اتصال جداگانه TCP به سرویس‌دهندهٔ موردنظر برقرار شود [۱۲] (شکل ۳-ب).



(الف)

[1] Object
[2] Resource



(ب)

(شکل-۳): (الف) ساختار یک صفحه وب – (ب) فرآیند بارگذاری یک صفحه وب

۲-۱- پروتکل HTTP

پروتکل HTTP یک پروتکل بدون حالت[3] است که هنگام مرور وب، برای تبادل اطلاعات بین کاربر و سرویس‌دهنده، به کار می‌رود [۱۳]. تاکنون نسخه‌های گوناگونی از پروتکل HTTP ارائه شده که در حال حاضر نسخهٔ HTTP/1.1 دارای بیشترین میزان استفاده در میان کاربران است[4]. تفاوت اصلی پروتکل HTTP/1.1 با نسخه قدیمی خود (HTTP/1.0) در وجود اتصالات پایدار[5] برای دریافت صفحات وب است. در پروتکل HTTP/1.0 برای دریافت هر منبع، لازم است یک اتصال TCP مجزا برقرار شود و پس از دریافت کامل آن منبع، ارتباط بسته‌می‌شود، درحالی‌که در پروتکل HTTP/1.1 امکان دریافت منابع موجود در یک سرویس‌دهنده با یک اتصال TCP وجود دارد؛ هم‌چنین در این نسخه از پروتکل، برقراری چندین اتصال موازی به چندین سرویس‌دهنده برای دریافت منابع مختلف امکان‌پذیر است (هرچند تعداد اتصالات موازی محدود و وابسته به نوع مرورگر است) [۱۵].

در یک اتصال TCP در هر لحظه، تنها ارسال یک درخواست امکان‌پذیر است. درخواست‌های بعدی، تنها زمانی ارسال می‌شوند که پاسخ درخواست پیشین به طور کامل دریافت شده‌باشد. یکی دیگر از ویژگی‌های پروتکل HTTP/1.1 فعال‌سازی خط لولهٔ HTTP[6] است [۱۲]. این ویژگی به مرورگر اجازه می‌دهد چندین درخواست GET را به طور هم‌زمان و به‌ترتیب ارسال کند و درصورتی‌که سرویس‌دهنده از این امکان پشتیبانی کند، می‌تواند به درخواست‌های مرورگر به‌ترتیب پاسخ دهد.

[3] Stateless
[4] بر اساس گزارش w3techs [۱۴] درحال حاضر ۸۶/۴ درصد از کل وب سایت‌های دنیا از پروتکل HTTP/1.1 استفاده می‌کنند.
[5] Persistent
[6] HTTP Pipelining

۳- حملات انگشت‌نگاری صفحات وب

در روش‌های انگشت‌نگاری تارنما، برای هر صفحه وب تحت‌نظارت مهاجم براساس نمونه‌های آموزشی آن صفحه، به صورت صریح یا ضمنی یک نمایهٔ[1] جدید ایجاد می‌شود. برای شناسایی هر دنبالهٔ بسته جدید، این نمایه به همراه نمایهٔ ایجادشده از نمونهٔ جدید، به‌عنوان ورودی الگوریتم یادگیری ماشین یا یک روش آماری استفاده می‌شود.

در این بخش، حملات انگشت‌نگاری، به تفکیک ویژگی‌های مورداستفاده و نمایه‌های ایجادشده برای هر صفحه وب در هفت زیربخش، بررسی می‌شود؛ هم‌چنین سناریو، نحوهٔ ارزیابی حملهٔ ارائه‌شده در هر پژوهش و سایر جزئیات مربوط به آن، در صورت لزوم بیان می‌شود. ناگفته نماند که فرض نخستین تمامی این روش‌ها انتقال ترافیک کاربر به صورت رمزشده است.

در جدول‌های (۱ تا ۳) و جدول‌های (۶ و ۷) روش‌های معرفی‌شده در این بخش، مجموعهٔ داده‌های مورداستفاده و نتایج حاصل از ارزیابی هر روش، بر حسب دقت[2] گزارش‌شده در آن پژوهش، به‌اختصار آمده‌است. یادآوری می‌شود که پارامترهای مربوط به پیاده‌سازی هر روش، از جمله اندازهٔ مجموعه‌های آموزش و آزمون و مجموعهٔ دادهٔ استفاده‌شده در هر پژوهش، متفاوت است؛ هم‌چنین به‌دلیل نبود اطلاعات کافی دربارهٔ سرعت اجرای برخی از روش‌های معرفی‌شده در پژوهش‌های بالا، این شاخصه در جدول ذکر نشده‌است.

۳-۱- حملات مبتنی بر اندازهٔ شیء

در نخستین حملات شناسایی ترافیک رمزشده، از اندازهٔ اشیای موجود در هر صفحه برای شناسایی صفحه استفاده می‌شد [۱۶-۱۹] (جدول ۱).

با ظهور پروتکل SSL3.0، تلاش‌هایی برای بررسی میزان امنیت و میزان نشت اطلاعات در انتقال ترافیک با استفاده از این پروتکل انجام شد. در یکی از نخستین بررسی‌های انجام‌شده در تحلیل پروتکل SSL، بیان شد که شناسایی ترافیک یک صفحه وبِ مشاهده‌شده توسط کاربر با تحلیل طول درخواست GET رمزشده توسط پروتکل SSL امکان‌پذیر است [۱۶]. در روش استفاده‌شده چنگ و آونور [۱۷]، برای صفحات وب گوناگون، نمایه‌ای شامل جفت اندازهٔ کد HTML، اندازه شیء ایجاد و در یک پایگاه داده ذخیره می‌شود، سپس، با در نظر گرفتن سه صفحه وب پی‌درپی ملاقات‌شده توسط کاربر و رجوع به پایگاه داده، احتمال بازدید یک صفحه وب به کمک الگوریتم تحلیل لینک محاسبه می‌شود و صفحه با امتیاز بیشتر به‌عنوان صفحهٔ ملاقات‌شده توسط کاربر اعلام می‌شود. در این الگوریتم، یک صفحهٔ وب از پایگاه داده به طور تصادفی انتخاب می‌شود؛ سپس، برای نمونه ورودی جدید، احتمال این که صفحهٔ موردنظر، صفحه p باشد، مطابق رابطه (۱) محاسبه می‌شود و بالاترین امتیاز، مشخص‌کنندهٔ صفحه وب نمونه جدید است. در این رابطه، p^{-1} معادل احتمال بازدید صفحه p بلافاصله پس از بازدید صفحه p^{-1} و p^{+1} احتمال بازدید صفحه p^{+1} بی‌درنگ پس از بازدید p است (بنابراین این روش نیاز به پایگاه دادهٔ بزرگی از رفتارهای کاربر در هنگام بازدید صفحات وب دارد).

$$P(page) = P(p^{-1}).P(p^{+1}) \quad (۱)$$

هینتز [۱۸] نشان داد که با در نظر گرفتن اندازهٔ اشیای موجود در یک صفحه و تعداد دفعات تکرار این اندازه، برای هر صفحه یک نمایه یکتا می‌توان ایجاد کرد؛ سپس، با مقایسهٔ تعداد تطبیق‌های موجود بین نمایهٔ یک صفحه وب مشخص با نمایهٔ نمونهٔ جدید، می‌توان دربارهٔ نوع ترافیک نمونهٔ جدید تصمیم‌گیری کرد. در این پژوهش، شناسایی با این فرض انجام شد که کاربران از پراکسی تک‌گام SafeWeb استفاده می‌کنند و ارزیابی روش پیشنهادی برای شناسایی پنج صفحهٔ وب خبری مشهور نشان داد که حتی با وجود پراکسی و داده‌های رمزشده نیز امکان شناسایی صفحات وب وجود دارد؛ به‌طور مشابه، سان و همکاران [۱۹] با در نظر گرفتن تعداد و اندازهٔ اشیای واکشی‌شده توسط مرورگر و انتخاب معیار مشابهت ضریب جاکارد[3]، شناسایی یک صفحه وب از میان مجموعه‌ای از صفحات هدف را بررسی کردند.

در پژوهش‌های یادشده، استفاده از ترافیک TCP ناپایدار[4] برای تبادل داده بین مشتری و سرویس‌دهنده، امکان دسترسی به تعداد و سایز اشیا را فراهم می‌کرد. در حال حاضر، با فراگیرشدن استفاده از پروتکل HTTP/1.1 دسترسی به اندازه شیء‌های موجود در صفحه امکان‌پذیر نیست.

[1] Profile
[2] Accuracy
[3] Jaccard coefficient
[4] Non-persistent TCP

(جدول-۱): حملات انگشت‌نگاری صفحات وب معرفی‌شده در بخش‌های ۳-۱ تا ۳-۳ - دقت ثبت‌شده در جدول، برحسب درصد و مربوط به نتایج به‌دست‌آمده از هر پژوهش با در نظر گرفتن مجموعهٔ داده و نوع ارزیابی گفته‌شده در ستون ارزیابی جدول است.

ردیف	نویسنده	روش دسته‌بندی	مجموعه ویژگی‌ها	مجموعه داده	ارزیابی	دقت (٪)
۱	چنگ و آونور (۱۹۹۸)[۱۷]	محاسبهٔ احتمال ملاقات یک صفحه بر اساس صفحات قبلی و بعدی در یک تارنما	اندازهٔ کد HTML، اندازهٔ شیء	صفحات وب سه تارنمای Enough[1]، Unex[2] و Spectator[3] بارگذاری‌شده با پروتکل HTTP		
۲	هینتز (۲۰۰۳)[۱۸]	مقایسهٔ تعداد تطبیق در نمونهٔ جدید و نمونهٔ موجود در پایگاه داده	اندازه اشیای صفحه، تعداد دفعات تکرار هر اندازه	صفحات وب پنج تارنمای خبری CNN[4]، BBC[5]، نیویورک تایمز[6]، اسلش دات[7] و واشینگتن پست[8] بارگذاری شده با SafeWeb	جهان بسته	گزارش نشده
۳	سان و همکاران (۲۰۰۲)[۱۹]	ضریب جاکارد	تعداد و اندازه شیءهای واکشی‌شده توسط مرورگر	صفحات بارگذاری شده از پایگاه داده DMOZ[9]		
۴	شماتیکوف و وانگ(۲۰۰۶)[۲۰]	آزمون هم‌بستگی	تعداد بسته‌های پنجره‌های زمانی به طول W در دنباله	شبیه‌سازی ترافیک با استفاده از مجموعه دادهٔ NLANR[10] شامل دنباله‌های ترافیک HTTP	جهان بسته	گزارش نشده
۵	بیسیاس و همکاران(۲۰۰۵)[۲۳]	آزمون هم‌بستگی	دنبالهٔ طول بسته، دنبالهٔ زمان بین ورود	ترافیک HTTP تونل‌شده با OpenSSH [۲۴]		

۳-۲- حملات مبتنی بر تعداد بسته‌های دنباله

در پژوهش شماتیکوف و وانگ [۲۰] شبکه‌های mix (مانند شبکه‌های معرفی‌شده در پژوهش دینگلداین و همکاران [۲۱] و پژوهش فریدمن و موریس [۲۲])، به‌عنوان بستر تبادل اطلاعات کاربران به‌منظور پنهان کردن مبدأ و مقصد ارتباط در نظر گرفته‌شدند و روشی برای شناسایی صفحات وب مرورشده توسط کاربر ارائه شد. در این روش، مهاجم در بازهٔ زمانی شصت ثانیه به مشاهدهٔ ترافیک می‌پردازد؛ سپس این بازه را به پنجره‌های زمانی با طول W تقسیم می‌کند. با شمارش تعداد بسته‌های موجود در هر پنجره، یک دنبالهٔ عددی برای هر ترافیک ایجاد می‌شود (P_i). به‌منظور شناسایی ترافیک، هم‌بستگی[11] بین P_i و دنباله‌های P_c موجود در مجموعهٔ آموزشی ($\chi(P_i, P_c)$) مطابق با رابطهٔ (۲) محاسبه شده و درصورتی‌که عدد به‌دست‌آمده از یک حد آستانه (τ) بیشتر باشد (طبق رابطه (۳))، دو دنباله مشابه یکدیگر تشخیص داده‌می‌شوند؛ (به عبارت دیگر، مقصد ترافیک هر دنباله یکسان در نظر گرفته‌می‌شود).

$$\chi(P_i, P_c) = \frac{\sum_j (P_{i,j} - \mu_{P_i})(P_{c,j} - \mu_{P_c})}{\sigma_{P_i}\sigma_{P_c}}, \quad (\sigma_Q = \sqrt{\frac{1}{|Q|}\sum_{j=1}^{|Q|}(Q_j - \bar{Q})^2}, \quad \mu_Q = \frac{\sum_{j=1}^{|Q|} Q_j}{|Q|}) \qquad (۲)$$

$$\ell(P_i) = argmax_{c\in T}(\chi(P_i, P_c) > \tau) \qquad (۳)$$

۳-۳- حملات مبتنی بر زمان بین ورود و اندازهٔ بسته‌ها

بیسیاس و همکاران [۲۳] دنباله‌های رمزشده HTTP و به طور ویژه جریان‌های رمزشده با پروتکل WEP/WPA، IPSec یا تونل‌های SSH را (که در آن‌ها نشانی مقصد جریانِ داده به دلیل رمزشدن سرایند بسته‌های IP در دسترس نیست)، مدنظر قرار دادند. در این پژوهش، دنباله‌های صد صفحه وب در طول مدت دو ماه جمع‌آوری شد (این مجموعه با نام WebIdent در پایگاهِ دادهٔ شبکهٔ دانشگاه ماساچوست آمرست قابل‌دسترسی است [۲۴]). برای هر دنباله، طول بسته‌ها و

[1] www.enough.org
[2] http://www.unex.berkeley.edu:4243
[3] www.spectator.org
[4] www.cnn.com
[5] www.bbc.co.uk
[6] www.nytimes.com
[7] slashdot.org
[8] www.washingtonpost.com
[9] DMOZ Open Directory Project link database (DMOZ.org)
[10] National Laboratory for Applied Network Research (pma.nlanr.net)
[11] Cross-correlation

زمان بین ورود آن‌ها به صورت $I.T = \{t_1, t_2, \dots t_n\}$ و $P.S = \{s_1, s_2, \dots s_n\}$ استخراج شد. به‌منظور ایجاد نمایهٔ هر صفحهٔ وب، برای هر دو دنباله زمان بین ورود و اندازهٔ متوسط همهٔ مقادیر $x_i (i \in [1,n])$ محاسبه و دنباله‌های جدید $\widehat{I.T}$ و $\widehat{P.S}$ با این مقادیر ایجاد شد. به‌منظور شناسایی مقصد ترافیک ورودی جدید (P_i)، هم‌بستگی بین دنبالهٔ جدید و نمایه ایجادشده برای هر صفحهٔ وب به‌ازای هر دو دنبالهٔ اندازه و زمان بین ورود اندازه‌گیری می‌شود و حاصل ضرب دو مقدار محاسبه می‌شود (رابطه (۴)).

$$\ell(P_i) = argmax_{c\in T}(\chi(P.S_i, \widehat{P.S_c})\chi(I.T_i, \widehat{I.T_c})) \quad (۴)$$

عدد به‌دست‌آمده به‌عنوان امتیاز شباهت صفحهٔ وب تازه و نمایه صفحات وب موجود در مجموعهٔ آموزشی در نظر گرفته‌می‌شود. بالاترین امتیاز به‌دست‌آمده به‌عنوان پیش‌بینی مقصد ترافیک جدید انتخاب می‌شود.

۴-۳- حملات مبتنی بر اندازه و جهت بسته‌های دنباله

نخستین روش شناسایی مقصد ترافیک بر اساس اندازهٔ بسته، توسط میستری و رامان [۲۵] و بر روی چهار سایت و صفحات آن‌ها انجام شد و نشان داده‌شد که ترافیک SSL به‌تنهایی قابلیت حفظ محرمانگی اطلاعات کاربر را ندارد. این کار با مقایسهٔ طول بسته‌های دریافت‌شده از سرویس‌دهنده با طول بسته‌های یک نمونه ترافیک جدید انجام گرفت.

در روش ارائه‌شده توسط لیبراتور و لوین [۲۶] به‌ازای هر صفحه وب موجود در مجموعهٔ تحت‌نظارت مهاجم، مجموعه‌ای از دوتایی‌های جهت-اندازه تشکیل می‌شود. درصورتی‌که دوتایی جهت-اندازه با مقدار مشخص در بیشینهٔ داده‌های آموزشی متعلق به یک صفحه، موجود باشد، این دوتایی به نمایهٔ صفحه وب موردنظر مهاجم اضافه می‌شود؛ به عبارت دیگر، در این روش، بردار ویژگی یا نمایهٔ هر طبقه (هر صفحه وب)، فهرستی از اندازهٔ بسته‌هاست. به منظور شناسایی مقصد ترافیک ورودی (صفحه وب بازدیدشده)، میزان شباهت ترافیک جدید با هرکدام از نمایه‌های از پیش جمع‌آوری‌شده با ضریب جاکارد (رابطه (۵)) محاسبه می‌شود. ترافیک جدید، مربوط به صفحه وب با بالاترین امتیاز شباهت خواهدبود (مجموعهٔ نمایهٔ ایجادشده برای یک صفحه وب را X و مقادیر جمع‌آوری‌شده برای ترافیک ورودی را Y می‌نامیم).

$$\ell(P_i) = argmax_{c\in T} S(X_c, Y) = argmax_{c\in T} \frac{|X_c \cap Y|}{|X_c \cup Y|} \quad (۵)$$

در همین پژوهش [۲۶] ، از دسته‌بند بیز ساده[1] برای دسته‌بندی ترافیک جدید استفاده شد. در این روش هر اندازه یکتای بسته (با در نظر گرفتن جهت آن) یک ویژگی محسوب می‌شود و مقدار هر ویژگی برابر با تعداد تکرار اندازه بسته موردنظر است. مجموعه ویژگی به‌دست‌آمده به‌صورت $P.S$ خواهدبود.

$$P.S = \{(s_1, cnt_{s_1}), (s_2, cnt_{s_2}), \dots, (s_n, cnt_{s_n})\}$$

برای شناسایی یک صفحه وب تازه، احتمال تعلق نمونه P_i به کلاس C به‌ازای کلیهٔ صفحات وب تحت‌نظارت محاسبه می‌شود (رابطهٔ (۶)). واضح است که در این روش، دقت دسته‌بندی وابسته به تعداد دقیق دفعات تکرار طول بسته‌ها (مشخصهٔ cnt_{s_i}) است.

$$p(P_i \in C) = p(C|P.S) = \frac{P(C)P(P.S|C)}{P(P.S)} \propto p(C)\prod_{j=1}^{n} p(cnt_{s_j}|C) \quad (۶)$$

در پژوهش یادشده، ترافیک HTTP رمزشده در تونل‌های چندگام (مثل Tor و JAP)، تونل‌های SSH و IPSec و اتصالات رمزشده توسط پروتکل WPA موردنظر قرار گرفته‌است و در پیاده‌سازی از OpenSSH و پراکسی SOCKS استفاده شده‌است[۲۷].

به طور مشابه، در پژوهش هرمن و همکاران [۶]، شناسایی ترافیک صفحات وب برای شبکه‌های ناشناس Tor، JAP و هم‌چنین بر روی OpenSSH، OpenVPN و STunnel با استفاده از روش بیز ساده چندجمله‌ای پیاده شد. روش بیز چندجمله‌ای به‌طورمعمول برای دسته‌بندی متن استفاده می‌شود. یک دنبالهٔ بسته را با همه اندازه‌های بستهٔ مشاهده‌شده در آن می‌توان مشابه یک متن شامل تعدادی کلمه با تعداد تکرارهای گوناگون در نظر گرفت. در این رویکرد، $P(P.S|C)$ در فرمول بیز ساده مطابق رابطه (۷) محاسبه می‌شود.

[1] Naïve Bayes

(جدول-۲): حملات انگشت‌نگاری صفحات وب مبتنی بر اندازه و جهت بسته‌های دنباله – دقت ثبت شده‌در جدول برحسب درصد و مربوط به نتایج به‌دست‌آمده از هر پژوهش با در نظر گرفتن مجموعه داده و نوع ارزیابی گفته‌شده در ستون ارزیابی جدول است.

ردیف	نویسنده	روش دسته‌بندی	مجموعه ویژگی‌ها	مجموعه داده	ارزیابی	دقت(٪)
۱	میستری و رامان(۱۹۹۸) [۲۵]	مقایسهٔ نمونهٔ جدید با نمونه‌های موجود در پایگاه داده	طول بسته‌های ورودی	ترافیک HTTPS صفحات وب چهار تارنمای ISSAC[1]، INDEX[2]، Wells Fargo[3] و Datek[4]	-	ناموجود
۲	لیبراتور و لوین (۲۰۰۶) [۲۶]	ضریب جاکارد	مجموعهٔ دوتایی‌های اندازه-جهت	ترافیک جمع‌آوری‌شده HTTP تونل‌شده با OpenSSH [۲۷]	جهان بسته	۸۶٪
۳	لیبراتور و لوین (۲۰۰۶) [۲۶]	بیز ساده	(اندازه، جهت) بسته، تعداد تکرار هر (اندازه، جهت)	ترافیک جمع‌آوری‌شده HTTP تونل‌شده با OpenSSH [۲۷]	جهان بسته	۸۳٪
۴	هرمن و همکاران(۲۰۰۹) [۶]	بیز چندجمله‌ای	همهٔ اندازه-جهت‌های مشاهده‌شده در دنباله صفحات گوناگون	ترافیک CiscoVPN، OpenSSH، STunnel و OpenVPN و ترافیک شبکه چندگام Tor	جهان بسته	۹۶/۶۵٪ (OpenSSH) ۲/۹۶٪ (Tor)

$$P(P.S|C) \propto \prod_{j=1}^{m} p(s_j|C)^{cnt_{s_j}} \quad (۷)$$

ارزیابی‌های این پژوهش نشان داد که روش ارائه‌شده برای شناسایی صفحات وب در سامانه‌های تک‌گام مؤثر است، امّا برای شبکه‌های پراکسی چندگام Tor و JAP ناموفّق عمل می‌کند. حملات معرفی‌شده در این بخش در جدول (۲) نمایش داده شده‌است.

۵-۳- حملات مبتنی بر ترتیب بسته‌های دنباله

در پژوهش‌های بیان‌شده در بخش‌های پیش، هر نمایه با استفاده از ویژگی‌های یادشده و صرفاً به صورت مجموعه‌ای از اعداد تشکیل می‌شد. هینتز [۱۸] به‌منظور بهبود حملهٔ انگشت‌نگاری صفحات وب، ترتیب ارسال درخواست و دریافت اشیا را بررسی کرد و نشان داد که در مرورگرهای مختلف، این ترتیب متفاوت است و برای هر مرورگر ویژه، ترتیب بسته‌ها ثابت است. اگرچه، همان‌طور که پیشتر گفته‌شد در این پژوهش فرض ناپایداربودن اتصالات TCP در نظر گرفته شده است.

مشابه روش‌های پیشین، لو و همکاران [۲۸] برای شناسایی صفحه وب، ترافیک HTTP رمزشده را که توسط VPN یا SSH تونل شده‌است، انتخاب کردند و به این منظور، از دو مجموعهٔ داده شامل دنباله‌های بستهٔ بارگذاری صفحات توسط OpenSSH و OpenVPN استفاده کردند. در این روش از پروتکل HTTP پایدار برای بارگذاری صفحات استفاده شد. دنبالهٔ طول بسته‌ها (P) با در نظر گرفتن ترتیب آن‌ها به دو دنباله شامل بسته‌های درخواست ($P^{Request}$) و بسته‌های پاسخ ($P^{Response}$) تقسیم شد. بسته‌هایی که در دنبالهٔ پاسخ، طولی برابر با حداکثر اندازه قابل انتقال (MTU[5]) داشتند، حذف شدند؛ هم‌چنین، بسته‌هایی با طول کمتر از یک حد

[1] www.issac.cs.berkeley.edu
[2] www.index.berkeley.edu
[3] Banking.wellsfargo.com
[4] Orders.datek.com
[5] Maximum Transfer Unit

آستانه به‌عنوان بسته‌های کنترلی جریان در نظر گرفته‌شده و حذف شدند. به این ترتیب، برای هر صفحه وب موردنظر، یک نمایه شامل دو دنباله بالا ایجاد شد. برای شناسایی هر صفحه ورودی جدید، از فاصلهٔ ویرایش لونشتاین[1] استفاده شد. فاصله لونشتاین تعداد عملیات درج[2]، حذف[3] و جایگذاری[4] لازم را برای تبدیل یک دنباله به دنبالهٔ دیگر نشان می‌دهد. فاصله نمایهٔ صفحات وب با نمونه جدید از رابطهٔ (۸) تعیین می‌شود. کمترین فاصله، نشان‌دهندهٔ صفحه وب موردنظر خواهدبود.

$$d(P_i, P_c) = \alpha \cdot d_{lev}(P_i^{Request}, P_c^{Request}) + (1-\alpha) \cdot d_{lev}(P_i^{Response}, P_c^{Response}), \quad \alpha = 0.6 \tag{۸}$$

در پژوهش کای و همکاران [۲۹] و همچنین در پژوهش وانگ و گُلدبِرگ [۱]، هر دنبالهٔ ترافیک به‌صورت رشته‌ای از اعداد صحیح شامل طول بسته‌های دنباله (با حفظ ترتیب) در نظر گرفته‌شد و برای احتساب جهت هر بسته، علامت مثبت برای بسته‌های خروجی و علامت منفی برای بسته‌های ورودی تعیین شد؛ هم‌چنین در هر دو روش، شناسایی صفحهٔ وب با استفاده از دسته‌بند SVM انجام گرفت، با این تفاوت که برای محاسبهٔ فاصله بین نقاط داده (دنباله‌های ترافیک) از دو معیار جدید استفاده شد که به‌طورعموم در مسائل متن‌کاوی و برای تطبیق کلمات به کار می‌روند. در روش کای و همکاران [۲۹]، فاصلهٔ بهینهٔ تطبیق رشته (OSAD[5]) (با افزودن عملیات انتقال[6] به مجموعه عملیات لونشتاین) استفاده شد. هرچند به عملیّات انتقال هزینه کمتری نسبت به عملیات درج و حذف اختصاص یافت. در روش وانگ و گلدبرگ [۱] از فاصلهٔ دامرا-لونشتاین (DLD[7]) برای محاسبه فاصله بین دنباله‌های ترافیک استفاده شد. در این پژوهش برای دستیابی به دقت مناسب، عملیات جای‌گذاری از مجموعه عملیّات ممکن در روش لونشتاین حذف شد؛ هم‌چنین، هزینه عملیّات ویرایش برای بسته‌های خروجی و ورودی و برای ابتدا و انتهای دنباله، متفاوت در نظر گرفته شد. خلاصهٔ چهار حملهٔ بالا در جدول (۳) آمده‌است.

[1] Leveneshtein Edit Distance
[2] Insertion
[3] Deletion
[4] Substitution
[5] Optimal String Alignment Distance
[6] Transposition
[7] Damerau-Levenshtein distance

۶-۳- حملات مبتنی بر مجموعهٔ ویژگی‌ها

از سال ۲۰۱۱ ویژگی‌های انتخابی برای شناسایی تارنما کامل‌تر شدند و جزئیات بیشتری از فرآیند بارگذاری صفحات در نظر گرفته‌شد. پژوهش پانچنکو و همکاران [۳۰] نخستین پژوهشی بود که در آن شناسایی ترافیک در بستر شبکه‌های گم‌نامی از جمله شبکهٔ Tor و JAP بررسی و به آن بسیار توجه شد. در این اثر، همچنین مفهوم نشان‌گر اندازه[8] (به معنای محل تغییر جهت بسته‌ها در یک دنباله)، برای نخستین‌بار معرفی و از این ویژگی برای دسته‌بندی استفاده شد؛ هم‌چنین، با حذف بسته‌هایی با سایز ۵۲ بایت در دنباله به‌عنوان بسته‌های تصدیق دریافت (ACK)، دقت مطلوبی در شناسایی ترافیک در حضور شبکه Tor به‌دست آمد. ویژگی‌های مورداستفاده در این پژوهش مطابق جدول (۴) است.

در این پژوهش از ماشین بردار پشتیبان (SVM[9]) به منظور دسته‌بندی و شناسایی ترافیک تولیدشده توسط شبکه‌های Tor و JAP استفاده شد.

ایدهٔ اصلی SVM تبدیل نمونه‌ها به مجموعه‌ای از بردارهاست. با بارگذاری هر صفحه وب، داده‌های خام (دنباله بسته‌های ترافیک) به‌ازای هر صفحه به دست می‌آید و مجموعهٔ ویژگی‌ها از هر دنباله استخراج می‌شود. دسته‌بند تلاش می‌کند تا بر اساس دادهٔ آموزشی، یک ابرصفحهٔ جداکننده[10] را به فضای برداری نگاشت کند، به‌صورتی‌که فاصلهٔ نمونه‌های آموزشی موجود در دسته‌های گوناگون در راستای عمود بر این ابرصفحه بیشترین مقدار ممکن باشد. اگر حاشیهٔ[11] یک ابرصفحه به‌صورت مجموع فواصل نزدیک‌ترین نقاط آموزشی هر کلاس به ابرصفحه (بردارهای پشتیبان هر کلاس) تعریف کنیم، ابرصفحهٔ جداکننده یا مرز تصمیم‌گیری مناسب برای دسته‌بندی، ابرصفحه‌ای با بیش‌ترین مقدار حاشیه خواهدبود [۳۱]. درصورتی‌که بردارها به صورت خطی قابل جداسازی نباشند، فضای برداری به فضایی با ابعاد بیشتر تبدیل می‌شود و بار دیگر یک ابرصفحه برای نگاشت جستجو می‌شود [۳۰].

دایر و همکاران [۳۲] تلاش کردند با در نظر گرفتن سه دسته از ویژگی‌های کلی ترافیک (صرف‌نظر از ویژگی‌هایی مانند طول بسته‌های منفرد موجود در دنباله که در

[8] Size marker
[9] Support Vector Machine
[10] Separating Hyperplane
[11] Margin

(جدول-۳): حملات انگشت‌نگاری صفحات وب مبتنی بر ترتیب بسته‌های دنباله – دقت ثبت‌شده در جدول برحسب درصد و مربوط به نتایج به دست آمده از هر پژوهش با در نظر گرفتن مجموعه داده و نوع ارزیابی گفته‌شده در ستون ارزیابی جدول است.

ردیف	نویسنده	روش دسته‌بندی	مجموعه ویژگی‌ها	مجموعه داده	ارزیابی	دقت(٪)
۱	هینتز(۲۰۰۳) [۱۸]	مقایسه تعداد تطبیق در نمونه جدید و نمونه موجود در پایگاه داده	دنباله اندازه اشیاء و تعداد تکرار هر اندازه (با در نظر گرفتن ترتیب)	مشابه ردیف ۲ جدول	جهان بسته	گزارش نشده
۲	لو و همکاران(۲۰۱۰) [۲۸]	فاصله لونشتاین	دنباله طول بسته‌های ورودی، دنباله طول بسته‌های خروجی	ترافیک HTTP تونل‌شده با OpenSSH [۲۷] و OpenVPN	جهان بسته	۸۱٪ (OpenSSH) ۹۷٪ (OpenVPN)
۳	کای و همکاران (۲۰۱۲) [۲۹]	محاسبه فاصله بین نمونه‌ها به روش OSAD و دسته‌بند SVM	رشته مرتب‌شده از اعداد صحیح شامل طول بسته‌های دنباله	ترافیک تولیدشده توسط مرورگر Tor [۳۸] و ترافیک تونل‌شده با OpenSSH	جهان بسته	۹۱/۶٪ (OpenSSH) ۸۳/۷٪ (Tor)
۴	وانگ و گلدبرگ (۲۰۱۳) [۱]	محاسبه فاصله بین نمونه‌ها به روش دامرا-لونشتاین و دسته‌بند SVM	رشته مرتب‌شده از اعداد صحیح شامل طول بسته‌های دنباله	ترافیک صفحات وب تولیدشده توسط مرورگر Tor [۳۸]	هردو	۹۱٪ (closed world) ۹۰٪ (open world)

پژوهش‌های پیشین استفاده می‌شد)؛ همچنین با کمک الگوریتم بیز ساده روشی با پیچیدگی کمتر ارائه دهند.

به این منظور، هرکدام از سه ویژگی «زمان کل ارتباط»، «پهنای باند در هر طرف ارتباط» و «الگوی بسته‌های متوالی هم‌جهت[1]» در ترافیک به صورت N-Gram با طول متغیر استفاده‌شدند. درواقع در این روش، یک N-Gram عبارت است از مجموعهٔ جهت-اندازه کلیه بسته‌های متوالی هم‌جهت که در دنباله موجودند. هم‌چنین این پژوهش‌گران با ترکیب این سه ویژگی، یک دسته‌بند دیگر ارائه کردند که ساده‌تر از روش پانچنکو و همکاران [۳۰] بود، امّا از لحاظ دقت نسبت به روش قبلی بهبود چشم‌گیری نداشت. در این پژوهش دسته‌بندی بر روی ترافیک HTTP رمزشده توسط SSH، ترافیک TLS و IPSec که در پژوهش‌های قبلی جمع‌آوری شده‌بود، انجام گرفت.

وانگ و همکاران [۹] در مقایسه با پژوهش‌های پیشین، بزرگ‌ترین مجموعهٔ ویژگی‌ها را برای شناسایی ترافیک صفحات وب به کار گرفتند که به‌طورتقریبی شامل ۴۰۰۰ ویژگی بود (جدول ۵). در این پژوهش جزئیات کامل دنباله ترافیک برای تعداد زیادی از بسته‌های ابتدایی جریان ترافیک استفاده شد که این امر اندازهٔ مجموعهٔ ویژگی‌ها را افزایش داده است. در این روش، دسته‌بندی با الگوریتم K- نزدیک‌ترین همسایهٔ (KNN[2]) انجام می‌شود و دسته‌بندی تنها زمانی نمونه جدید را به یک دسته منتسب می‌کند که هر K همسایه به آن دسته تعلق داشته باشد.

به منظور بهبود دقت شناسایی، روشی برای اختصاص وزن به مجموعهٔ ویژگی‌ها ارائه شد که طی آن وزن هر ویژگی (w_i) در یک الگوریتم تکرارشونده محاسبه می‌شود و سپس در محاسبه فاصله بین نمونه‌های آموزشی و نمونهٔ جدید اعمال می‌شود. رابطه (۹)، فاصله وزن‌دار بین نمونهٔ آموزشی P_i و نمونهٔ آزمون P_c را برای n ویژگی با مقادیر p نشان می‌دهد. استفاده از روش تخصیص وزن ویژگی‌ها به انتخاب ویژگی‌های مؤثر در دسته‌بندی و بهبود دقت دسته‌بند کمک می‌کند.

[1] Burst

[2] K-Nearest Neighbor

(جدول-۴): ویژگی‌های مورد استفاده در پژوهش پانچنکو و همکاران [۳۰]

ردیف	نام ویژگی	تعداد
۱	تعداد بسته‌های بین دو نشان‌گر اندازه	β (وابسته به تعداد نشان‌گرها)
۲	مجموع طول بسته‌های موجود بین دو نشان‌گر (بایت)	β (وابسته به تعداد نشان‌گرها)
۳	اندازه سند HTML (تعداد و اندازه بسته‌های دریافتی بین اولین بسته خروجی از سمت کاربر و بسته‌های خروجی بعدی)	۲
۴	مجموع اندازه بسته‌های ورودی و خروجی (بایت)	۲
۵	تعداد بسته‌های ورودی و خروجی	۲
۶	اندازه بسته‌های ورودی و خروجی (با احتساب جهت)	A
۷	تعداد تکرار اندازه بسته‌های موجود در هر نمونه	α
۸	درصد بسته‌های ورودی و خروجی	۲

$$d(P_i, P_c) = \sum_{j=1}^{n} w_j |p_{i,j} - p_{c,j}| \quad (۹)$$

در سال ۲۰۱۶، پانچنکو و همکاران [۱۰] برای استفاده از دنبالهٔ اندازه بسته‌ها، روش CUMUL را ارائه کردند. در این پژوهش به‌جای استفاده از طول بسته‌های دنباله به‌عنوان یک ویژگی مجزا و از پیش تعیین‌شده، از جمع تجمعی اندازه‌های بسته استفاده شده‌است که به‌صورت ضمنی همهٔ ویژگی‌های مرتبط با اندازهٔ بسته، از جمله محل قرارگیری بسته‌های متوالی هم‌جهت را پشتیبانی می‌کند؛ سپس از نمایش تجمعی به‌دست‌آمده نمونه‌برداری و از نقاط حاصل از نمونه‌برداری به‌عنوان ویژگی‌های دیگر دنباله برای بهبود دقت دسته‌بند استفاده شده است. این پژوهش، دقت دسته‌بند SVM خود را برای سه سطح مختلف سلول (در دنباله‌های تولیدشده توسط Tor، TLS و TCP) بررسی کرده است که تفاوت قابل‌توجهی نیز در دقت به‌دست‌آمدهٔ آن مشاهده نمی‌شود؛ هم‌چنین در این پژوهش، مقیاس‌پذیری روش شناسایی صفحه وب پیشنهادشده با شناسایی تارنما مقایسه و نشان داده‌شد که شناسایی ترافیک تارنما با افزایش اندازهٔ مجموعهٔ بدون نظارت، دقت پایدارتری دارد (در این پژوهش، برای مقایسهٔ تارنما و صفحه وب، ابتدا صفحهٔ خانهٔ تارنماهای موردنظر و سپس صفحات گوناگونی از هر تارنما بارگذاری و ترافیک آن‌ها جمع‌آوری شده‌است). آندری پانچنکو مجموعه دادهٔ جمع‌آوری‌شده توسط مرورگر تور را برای این پژوهش در صفحهٔ شخصی خود در قالب دو فایل برای داده‌های تحت‌نظارت و بدون نظارت به اشتراک گذاشته‌است [33].

(جدول-۵): ویژگی‌های مورد استفاده در پژوهش وانگ و همکاران [۹]

ردیف	نام ویژگی	تعداد
۱	تعداد کل بسته‌ها	۱
۲	تعداد بسته‌های ورودی و خروجی	۲
۳	زمان کل	۱
۴	اندازهٔ بسته‌های موجود در دنباله	۳۰۰۰ [-۱۵۰۰، ۱۵۰۰]
۵	مکان n بستهٔ خروجی ابتدایی	n =۳۰۰
۶	فاصلهٔ n بستهٔ خروجی ابتدایی با بستهٔ خروجی قبلی	n =۳۰۰
۷	بیشینهٔ تعداد بسته‌های خروجی در m پنجره با اندازه w بسته	m = ۱۰۰ (w = ۳۰)
۸	حداکثر طول burst در دنباله	۱
۹	متوسط طول burst های دنباله	۱
۱۰	تعداد burst های دنباله	۱
۱۱	تعداد burst های با طول بیشتر از l	l = {۵،۱۰،۱۵}
۱۲	طول b تا burst ابتدایی	b =۵
۱۳	جهت حرکت d بسته اول	d = ۲۰

در پژوهش هیز و دنسیز [۳۴]، روش KFP [1] معرفی شد. در KFP از جنگل تصمیم تصادفی [2] برای تولید نمایه برای هر دنباله بسته استفاده می‌شود. در این روش، خروجی هر درخت یک مؤلّفه از اثر انگشت یک نمونهٔ ورودی را مشخص می‌کند و مجموعهٔ این مؤلّفه‌ها بردار اثر انگشت نمونه را می‌سازد؛ سپس این اثر انگشت‌ها به الگوریتم KNN داده می‌شود. با محاسبهٔ فاصله همینگ بین دو اثر انگشت و رای‌گیری بین نزدیک‌ترین K نمونه آموزشی، می‌توان صفحه وب مرورشده توسط کاربر را شناسایی کرد.

[1] K-fingerprinting
[2] Random decision forest

(جدول-۶): حملات انگشت‌نگاری صفحات وب مبتنی بر مجموعه ویژگی‌ها – دقت ثبت شده در جدول برحسب درصد و مربوط به نتایج به‌دست‌آمده از هر پژوهش با در نظر گرفتن مجموعه داده و نوع ارزیابی گفته شده در ستون ارزیابی جدول است.

ردیف	نویسنده	روش دسته‌بندی	مجموعه ویژگی‌ها	مجموعه داده	ارزیابی	دقت(٪)
۱	پانچنکو و همکاران (۲۰۱۱) [۳۰]	دسته‌بند SVM	مطابق جدول (۲)	صفحات وب پیشنهادی پژوهش هرمن و همکاران [۶] بارگزاری شده توسط Tor و JAP	هرمن	۵۴/۶۱٪ (Tor closed world) ۷۶٪ (Tor open world)
۲	دایر و همکاران (۲۰۱۲) [۳۲]	بیز ساده	زمان کل ارتباط، پهنای باند در هر طرف ارتباط، الگوی burst ترافیک به صورت N-Gram با طول متغیر	مجموعه داده پژوهش هرمن و همکاران + مجموعه داده پژوهش لیبراتور و لوین[۲۷]	جهان-بسته	۹۰/۶٪* (مجموعه داده هرمن و همکاران)
۳	وانگ و همکاران [۹](۲۰۱۴)	الگوریتم تعیین وزن WLLCC و دسته‌بند KNN	مطابق جدول (۳)	ترافیک صفحات وب بارگزاری شده توسط شبکه چندگام Tor [۳۸]	جهان-باز	۷۹٪
۴	پانچنکو و همکاران (CUMUL)(۲۰۱۶) [۱۰]	دسته‌بند SVM	جمع تجمعی اندازه‌های بسته، نمونه‌برداری با تعداد متغیر از جمع تجمعی	ترافیک شبکه چندگام Tor [۳۳]	جهان-باز	۹۶٪

* در این پژوهش، دقت روش پانچنکو و همکاران [۳۰] در شرایط آزمایش برابر با ۹۶/۴ گزارش شده است.

در پژوهش النامی و همکاران [۲] با در نظر گرفتن دنباله‌های بسته‌های متوالی هم‌جهت مجاور که در دو جهت مخالف هستند و همچنین استفاده از اندازه و زمان آن‌ها، روشی برای شناسایی ارائه شد. برای دسته‌بندی صفحات وب در حالت جهان-بسته از ماشین‌های بردار پشتیبان و برای حالت جهان-باز از دسته‌بند KNN استفاده شد.

در این پژوهش به‌جای استفاده از الگوریتم تعیین وزن معرفی‌شده در پژوهش وانگ و همکاران [۹]، از جنگل تصمیم تصادفی برای تعیین وزن ویژگی‌ها استفاده شده‌است. این روش زمان محاسبات را نسبت به روش وانگ و همکاران کاهش می‌دهد.

پژوهش طائبی و همکاران [۳۵]، با در نظر گرفتن ماهیت پویای صفحات وب و لزوم به‌روزرسانی پیوسته مجموعه‌های آموزشی، روشی را برای افزایش سرعت آموزش دسته‌بند در حملات انگشت‌نگاری تارنما ارائه کردند. در این پژوهش، از یک دسته‌بند KNN وزن‌دار برای شناسایی ترافیک ورودی با بهره‌گیری از ویژگی‌های ارائه‌شده در پژوهش وانگ [۹] استفاده شد.

به‌منظـور کاهش ابعاد فضای ویژگی‌های نمونه‌های ترافیک، یک گام پیش‌پردازش داده به حملهٔ پیشنهادی اضافه شد. در این گام، مجموعهٔ بزرگ ویژگی‌های معرفی‌شده در پژوهش وانگ [۹] با استفاده از روش تحلیل مؤلّفه‌های اساسی (PCA[1]) به مجموعهٔ کوچکی با مقدار تقریبی ۲۳ ویژگی کاهش یافت که این امر باعث بهبود چشمگیر سرعت آموزش و آزمون دسته‌بند نیز شد. در پژوهش دیگری که توسط طائبی و همکاران [۳۶] انجام شد، تأثیر گام پیش‌پردازش داده با استفاده از تحلیل مؤلّفه‌های اساسی بر نرخ مثبت واقعی، نرخ مثبت کاذب، زمان آموزش و زمان آزمون دسته‌بند، با تغییر مشخصه‌های مؤثر در حمله برای دسته‌بند شبکهٔ عصبی بررسی و نشان داده‌شد که استفاده از PCA، ضمن کاهش هفده درصدی نرخ مثبت واقعی در ارزیابی جهان-باز، به افزایش بیش از پنجاه درصدی سرعت آموزش و آزمون می‌انجامد. در ارزیابی دو پژوهش بالا، از مجموعهٔ دادهٔ مورداستفاده در پژوهش وانگ و همکاران [۹] استفاده شد. در زمان پژوهش، این مجموعه داده، به‌روزترین و یکی از بزرگ‌ترین مجموعه‌های موجود از دنباله‌های بارگذاری‌شده از صفحات آغازین صد تارنما و شامل هجده هزار دنبالهٔ ترافیک

[1] Principal Component Analysis

است که در صفحهٔ شخصی وانگ به همراه تعدادی از مجموعه‌های داده مربوط به پژوهش‌های پیشین بارگذاری شده‌است[۳۷]. جداول (۶ و ۷) به حملات معرفی‌شده در این بخش و بخش بعد اختصاص دارد.

۷-۳- حملات مبتنی بر دنباله‌های خام

در یکی از آخرین پژوهش‌های منتشرشده در زمینه انگشت‌نگاری تارنما [۳۸]، به‌جای استفاده از یک تابع استخراج ویژگی برای انتخاب دقیق ویژگی‌ها، فرآیند استخراج ویژگی با به کارگیری شبکه‌های عصبی عمیق[1] به‌صورت خودکار انجام می‌شود. در این پژوهش، عملکرد سه شبکهٔ عصبی عمیقِ SDAE، CNN و LSTM در حوزهٔ انگشت‌نگاری تارنما برای تنظیمات جهان-بسته و جهان-باز بررسی شد. نتایج ارزیابی‌ها نشان داد که به طور کلی، دقت شناسایی این روش‌ها نسبت به دقیق‌ترین روش‌های این حوزه بهبودی نداشته، اما دقت آنها در حد این روش‌هاست. علاوه بر این، هرکدام از شبکه‌های عمیق مورداستفاده در شرایط ویژه، از لحاظ دقت یا سرعت یا میزان پایداری نتایج، نتایجی نزدیک به روش‌های قبلی داشته‌اند.

۴- نتیجه‌گیری

در این مقاله، حملهٔ انگشت‌نگاری صفحات وب به‌عنوان روشی برای کشف ماهیت ترافیک مرور وب کاربران معرفی شد و سیر تکامل پژوهش‌های انجام شده در این زمینه در طول زمان به شیوه‌ای نوین مورد بررسی قرار گرفت. به این منظور، پژوهش‌های پیشین انگشت‌نگاری تارنما بر اساس مجموعه ویژگی‌های استفاده‌شده برای شناسایی، به هفت دسته تقسیم و هرکدام از روش‌ها به اختصار توضیح داده‌شدند.

(جدول-۷): ادامهٔ حملات انگشت‌نگاری صفحات وب مبتنی بر مجموعهٔ ویژگی‌ها، دنباله‌های خام – دقت ثبت‌شده در جدول برحسب درصد و مربوط به نتایج به‌دست‌آمده از هر پژوهش با در نظر گرفتن مجموعهٔ داده و نوع ارزیابی گفته‌شده در ستون ارزیابی جدول است.

ردیف	نویسنده	روش دسته‌بندی	مجموعه ویژگی‌ها	مجموعه داده	ارزیابی	دقت(٪)
۱	هِیز و دِنسیز(۲۰۱۶) (KFP) [۳۴]	ایجاد اثرانگشت با استفاده از جنگل تصمیم تصادفی و دسته‌بندی با KNN	مطابق جدول (۳)	مجموعه دادهٔ پژوهش وانگ + ترافیک صفحات وب بارگذاری‌شده HTTPS و توسط Tor [۳۹]	جهان باز	۹۴٪
۲	النامی و همکاران (۲۰۱۶) [۲]	تعیین وزن ویژگی‌ها با جنگل تصمیم تصادفی و دسته‌بندی با KNN	اندازه بسته‌های ورودی و خروجی، اندازه و زمان Burst های تک‌جهتی و دوجهتی، تعداد بسته‌های موجود در Burst های تک‌جهتی	مجموعه دادهٔ پژوهش وانگ [۳۷] (Tor) + مجموعه دادهٔ پژوهش لیبراتور و لوین [۲۷] (HTTPS)	جهان باز	۹۲٪ (HTTPS) ۹۸٪ (Tor)
۳	طائبی و همکاران (۲۰۱۶) [۳۵]	دسته‌بند KNN	مؤلّفه‌های اساسی پراهمیت به‌دست‌آمده از مجموعه ویژگی‌های پژوهش وانگ	مجموعه دادهٔ پژوهش وانگ[۳۷]	جهان باز	۸۱٪
۴	طائبی و همکاران (۲۰۱۶) [۳۶]	دسته‌بند شبکهٔ عصبی	مؤلّفه‌های اساسی پراهمیت به دست‌آمده از مجموعه ویژگی‌های پژوهش وانگ	مجموعه دادهٔ پژوهش وانگ [۳۷]	جهان باز	۷۰٪
۵	ریمر و همکاران (۲۰۱۷) [۳۸]	شبکه‌های عصبی عمیق (LSTM،SDAE, CNN)	دنباله‌های ترافیک صفحات وب	ترافیک صفحات وب بارگذاری‌شده توسط Tor	جهان باز	۸۴٪ (شبکه SDAE)

[1] Deep Neural Networks

بررسی سیر تغییرات روش‌های انگشت‌نگاری نشان می‌دهد که به مرور زمان، در پژوهش‌های گوناگون، مجموعهٔ ویژگی‌های استفاده‌شده، جزئیات بیشتری از یک دنبالهٔ ترافیک را در بر گرفته‌اند که این امر موجب بزرگ‌ترشـدن مجموعهٔ ویژگی‌ها و صرف زمان بیشتر برای اجرای یک حمله می‌شود. با این حال، استفاده از ویژگی‌های دقیق‌تر، حملات ارائه‌شده را برای موفقیت در شناسایی صفحات وب در شرایط مختلف تقویت می‌کند و دقت شناسایی را افزایش می‌دهد.

۵- مراجع

[1] Wang, Tao and Goldberg, Ian, "Improved website fingerprinting on tor," in 12th ACM Workshop on privacy in the electronic society, 2013, pp. 201-212.

[2] Al-Naami, Khaled., Chandra, S., Mustafa, A., Khan, L., Lin, Z., Hamlen, K., Thuraisingham, B., "Adaptive encrypted traffic fingerprinting with bi-directional dependence," in Proceedings of the 32nd Annual Conference on Computer Security Applications, ACM, 2016, pp. 177-188.

[3] Aceto, Giuseppe and Pescapé, Antonio, "Internet Censorship detection: A survey," Computer Networks, 2015, vol. 83, pp. 381-421.

[4] Yanes, Adrian, "Privacy and Anonymity," arXiv preprint arXiv:1407.0423, 2014.

[5] Dixon, Lucas, Ristenpart, Thomas and Shrimpton, Thomas, "Network Traffic Obfuscation and Automated Internet Censorship," IEEE Security & Privacy, 2016, vol. 14, no. 6, pp. 43-53.

[6] Herrmann, Dominik, Wendolsky, Rolf and Federrath, Hannes, "Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier," in Proceedings of the 2009 ACM workshop on Cloud computing security, 2009, pp. 31-42.

[7] Cai, Xiang, Nithyanand, Rishab, Wang, Tao, Johnson, Rob and Goldberg, Ian, "A systematic approach to developing and evaluating website fingerprinting defenses," in Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014, pp. 227-238.

[8] Wang, Tao and Goldberg, Ian, "On realistically attacking Tor with website fingerprinting," in Pro-ceedings on Privacy Enhancing Technolog-ies, 2016, no. 4, pp.21-36.

[9] Wang, Tao, Cai, Xiang, Nithyanand, Rishab, Johnson, Rob and Goldberg, Ian, "Effective attacks and provable defenses for website fingerprinting," in 23rd USENIX Security Symposium (USENIX Security 14), 2014, pp. 143-157.

[10] Panchenko, Andriy., Lanze, F., Zinnen, A., Henze, M., Pennekamp, J., Wehrle, K. and Engel, T., "Website Fingerprinting at Internet Scale," in Proceedings of the 23rd Internet Society (ISOC) Network and Distributed System Security Symposium (NDSS 2016), 2016.

[11] Kurose, James. F. and Ross, K. W., Computer Networking: A Top-Down Approach, 6 ed., Pea-rson, 2012.

[12] Tanenbaum, Andrew S. and Wetherall, David J., Computer Networks, 5 ed., Prentice Hall, 2011.

[13] Fielding, R. and Reschke, J., "Internet Engineer-ing Task Force (IETF)," June 2014. [Online].

[14] Available: https://tools.ietf.org/html/rfc7230. [Accessed 25 August 2017].

[15] "Web Technology Surveys," [Online]. Available: https://w3techs.com/technologies/details/ce-http2/all/all. [Access-ed 25 August 2017].

[16] Nielsen, Henrik Frystyk, Gettys, James, Baird-Smith, Anselm, Prud'hommeaux, Eric, Wium Lie, Håkon and Lilley, Chris, "Network performance effects of HTTP/1.1, CSS1, and PNG." In ACM SIGCOMM Computer Communication Review, 1997, vol. 27, no. 4, pp. 155-166.

[17] Wagner, David and Schneier, Bruce, "Analysis of the SSL 3.0 protocol," in The Second USENIX Workshop on Electronic Commerce Proceedings, 1996, pp. 29-40.

[18] Cheng, Heyning and Avnur, Ron, "Traffic Analysis of SSL Encrypted Web Browsing," 1998. [Online]. Available: citeseer. ist. psu. edu/656522. html. [Accessed 15 September 2016].

[19] Hintz, Andrew, "Fingerprinting websites using traffic analysis," in International Workshop on Privacy Enhancing Technologies, Springer Berlin Heidelberg, 2002, pp. 171-178.

[20] Sun, Qixiang, Simon, D. R., Wang, Y., Russell, W., Padmanabhan, V. N. and Qiu, Lili, "Statist-ical identification of encrypted web browsing traffic," In Proceedings of 2002 IEEE Sympo-sium on Security and Privacy, 2002, pp. 19-30.

[21] Shmatikov, Vitaly and Wang, M.-H., "Timing analysis in low-latency mix networks:Attacks and defenses," in European Symposium on Research in Computer Security, Springer Berlin Heidelberg, 2006.

[22] Dingledine, Roger, Mathewson, Nick and Syverson, Paul, "Tor: The second-generation onion router," Naval Research Lab, Washington DC, 2004.

[23] Freedman, Michael J. and Morris, Robert, "Tarzan: A peer-to-peer anonymizing network

layer," in Proceedings of the 9th ACM conference on Computer and communications security, ACM, 2002, pp. 193-206.

[24] Bissias, George Dean, Liberatore, Marc, Jensen, D. and Levine, B. N., "Privacy vulnerabilities in encrypted HTTP streams," in International Workshop on Privacy Enhancing Technologies, Springer Berlin Heidelberg, 2005, pp. 1-11.

[25] WebIdent Traces, 2005, http://traces.cs.umass.edu/index.php/Network/Network, [Accessed 1 February 2017].

[26] Mistry, Shailen and Raman, Bhaskaran, "Quanti-fying Traffic Analysis of Encrypted Web-Browsing," 1998.

[27] Liberatore, Marc and Levine, Brian Neil, "Inferring the source of encrypted HTTP connec-tions," in Proceedings of the 13th ACM confer-ence on Computer and communications security. ACM, 2006, pp. 255-263.

[28] WebIdent2 Traces, 2006, http://traces.cs.umass.edu/index.php/Network/Network, [Accessed 1 February 2017].

[29] Lu, Liming, Chang, Ee-Chien and Chan, Mun, "Website fingerprinting and identification using ordered feature sequences," in Computer Security–ESORICS, 2010, pp. 199-214.

[30] Cai, Xiang, Zhang, Xin Cheng, Joshi, Brijesh, Johnson, Rob, "Touching from a Distance: Web-site Fingerprinting Attacks and Defenses," in 19th ACM Conference on Computer and Communications Security, 2012, pp. 605-616.

[31] Panchenko, Andriy, Niessen, Lukas and Zinnen, Andreas, "Website fingerprinting in onion routing based anonymization networks," in Proceedings of the 10th annual ACM workshop on Privacy in the electronic society, Chicago, IL, USA, 2011, pp. 103-114.

[32] Aggarwal, Charu C., Data mining: the textbook, Springer, 2015.

[33] Dyer, Kevin P., Coull, Scott E. and Ristenpart, Thomas, "Peek-a-boo, i still see you: Why efficient traffic analysis countermeasures fail," in 2012 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 2012, pp. 332-246.

[34] https://lorre.uni.lu/~andriy/zwiebelfreunde, [Accessed 1 February 2017].

[35] Hayes, Jamie and Danezis, George, "k-fingerprinting: a Robust ScalableWebsite Fingerprinting Technique," in Usenix Security Symposium, 2016, 1187-1203.

[36] طائبی، مریم، بهلولی، علی، کائدی، مرجان، "بهبود کارایی الگوریتم تشخیص حمله انگشت‌انگاری تارنما با استفاده از تحلیل مولفه‌های اساسی"، چهاردهمین دوره کنفرانس بین‌المللی انجمن رمز ایران، شیراز، شهریور ۱۳۹۶.

[37] طائبی، مریم، بهلولی، علی، کائدی، مرجان، " موازنه بین سرعت و دقت دسته‌بندی صفحات وب در انگشت‌نگاری صفحات وب "، نهمین دوره کنفرانس بین‌المللی فناوری اطلاعات و دانش (IKT2017)، تهران، مهر۱۳۹۶ .

[38] https://www.cse.ust.hk/~taow/wf/data/,[Accessed 1 February 2017].

[39] Rimmer, Vera, Preuveneers, Davy, Juarez, Marc, Goethem, Tom Van and Joosen, Wouter, "Automated Feature Extraction for Website Fingerprinting through Deep Learning," arXiv preprint arXiv:1708.06376, 2017.

[40] http://www.homepages.ucl.ac.uk/~ucabaye/alexa.tar.gz, [Accessed 1 February 2017].

مریم طائبی تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد مهندسی کامپیوتر به ترتیب در دانشگاه صنعتی اصفهان و دانشگاه اصفهان سپری کرده‌است. زمینهٔ پژوهشی موردعلاقه وی امنیت شبکه و یادگیری ماشین است.

علی بهلولی مقطع کارشناسی و کارشناسی‌ارشد خود را در دانشگاه صنعتی اصفهان سپری کرده و مدرک دکترای خود را از دانشگاه اصفهان گرفته است. وی از سال ۱۳۹۰ عضو هیئت علمی دانشکدهٔ مهندسی کامپیوتر دانشگاه اصفهان بوده و زمینه‌های پژوهشی وی شبکه و امنیت شبکه است.

مرجان کائدی دارای کارشناسی مهندسی کامپیوتر از دانشگاه صنعتی اصفهان و کارشناسی‌ارشد و دکتری مهندسی کامپیوتر از دانشگاه اصفهان است. وی از سال ۱۳۹۱ عضو هیئت علمی دانشکده مهندسی کامپیوتر اطلاعات دانشگاه اصفهان است. زمینه‌های پژوهشی مورد علاقهٔ وی، یادگیری ماشین، بهینه‌سازی و تجارت الکترونیک است.