

بررسی جنبه‌های امنیتی در مؤلفه‌ها و معماری

جویش‌گرهای بومی

حسن کوشککی^{۱*}، شفاقت نادری^۲، نسرین تاج نیشاپوری^۳ و مهسا امیدوار سرکندي^۴

^۱کارشناس ارشد امنیت مرکز تحقیقات مخابرات ایران، پژوهشکده امنیت، تهران، ایران
hassan.kooshkaki@gmail.com

^۲عضو هیأت علمی مرکز تحقیقات مخابرات ایران، پژوهشکده امنیت، تهران، ایران
naderi@itrc.ac.ir

^۳کارشناس ارشد امنیت مرکز تحقیقات مخابرات ایران، پژوهشکده امنیت، تهران، ایران
taj.ict@gmail.com

^۴کارشناس امنیت مرکز تحقیقات مخابرات ایران، پژوهشکده امنیت، تهران، ایران
mahsa.omidvarsarkandi@gmail.com

چکیده

امروزه با رشد شبکه‌های رایانه‌ای و افزایش فرازینده محتوا روی اینترنت، تقاضا برای تماشا و جستجوی محتواهایی مثل ویدیو، موسیقی، فایل‌ها و استناد زیاد شده است و جویش‌گرها یکی از عواملی هستند که در زمینه جستجو به هر تقاضایی از طرف کاربران پاسخ می‌دهند و به آنها کمک می‌کنند تا سریع‌تر به اهداف خود برسند. جویش‌گرها علاوه‌بر مزیت‌های بسیاری که دارند در صورت طراحی و پیکربندی ضعیف مؤلفه‌های خزش‌گر، نمایه‌ساز و رتبه‌بند، عملکرد نادرستی خواهند داشت و منجر به افسای اطلاعات محروم‌کاربران و وب‌گاه‌ها، ارائه نتایج غیر مرتبط و آلوه و غیر امن شدن معماري جویش‌گر می‌شوند؛ از این‌رو توجه به مسائل و جنبه‌های امنیتی مؤلفه‌ها در معماری جویش‌گرها، از جمله جویش‌گرهای بومی یکی از الزامات اساسی برای آنها محسوب می‌شود. مهم‌ترین بخش‌هایی که باید امنیت آنها در معماری جویش‌گرهای بومی تضمین شود، شامل مؤلفه خزش‌گر (سیاست‌های خزش‌گر، فاکتورهای طراحی، حملات تزریق کد، دسترسی‌پذیری خزش‌گرها و پایگاه داده خزش‌گرها)، مؤلفه نمایه‌ساز (واحد جستجو، پایگاه داده نمایه‌ساز، سازوکارهای نمایه‌سازی، فاکتورهای طراحی) و مؤلفه رتبه‌بند (سیاست‌های رتبه‌بندی، سیاست‌های دفاعی در برابر وب سایت‌های آلوه، بهینه‌سازی منفی جویش‌گرها یا سئو منفی یا به عبارت بهتر سئو کلاه‌سیاه) است. محورهای اصلی این مقاله در ابتدا به تهدیدها و آسیب‌پذیری‌های مؤلفه‌های اصلی جویش‌گرهای بومی پرداخته است؛ سپس در همین راستا الزامات و سیاست‌های امنیتی برای سه مؤلفه اصلی خزش‌گر، نمایه‌سازی و رتبه‌بند که منجر به امن‌سازی معماری جویش‌گرهای بومی می‌شود، ارائه شده است.

واژگان کلیدی: جویش‌گر، خزش‌گر، نمایه‌ساز، رتبه‌بند و سئو منفی.

۱- مقدمه

هر حمله تحت وب، تهدیدی برای آنها است. در ادامه ابتدا تهدیدهای مؤلفه‌های جویش‌گرها مشخص شده، سپس الزامات و سیاست‌های امنیتی که در جویش‌گرها باید وجود داشته باشد تا در برابر تهدیدها مقاوم باشند، تشریح شده است؛ زیرا وجود الزامات و سیاست‌های امنیتی برای مؤلفه‌های جویش‌گر، باعث امن‌شدن معماری جویش‌گر می‌شود.

رونده کار جویش‌گرها بدین صورت است که وقتی

امروزه با رشد شبکه‌های رایانه‌ای و پیشرفت در زمینه دستگاه‌هایی مثل تلفن‌های همراه هوشمند، تبلت‌ها، لپ‌تاپ‌ها و غیره تقاضا و جستجو برای محتواهای مبتنی بر وب افزایش یافته است. از طرفی با این رشد فرازینده در فناوری، تعداد حملات و آسیب‌پذیری‌ها نیز با همان ضریب افزایش پیدا کرده‌اند؛ به طوری که اثر و پیچیدگی حملات روی سامانه‌های قربانی نسبت به گذشته بیشتر شده است [1]؛ بنابراین از آنجاکه جویش‌گرها تحت وب عمل می‌کنند،

* نویسنده عهده‌دار مکاتبات

دو فصل نامه علمی ترویجی منادی امنیت فضای تولید و تبادل اطلاعات (افتا)

فرآیند فشرده‌سازی صفحات، آنها را در ذخیره Repository می‌کند و برای هر صفحه وب که در Repository ذخیره می‌شود، یک شماره ID به نام DocID تخصیص داده Doc index شود؛ DocIDها در واحدهای² Barrels و ذخیره می‌شوند.

Sorter و Indexer فرآیند نمایه‌سازی توسط مؤلفه‌های Sorter و Indexer انجام می‌شود. اسناد Indexer اسناد ذخیره شده در Repository را از حالت فشرده خارج و آنها را تجزیه می‌کند. هر سند به مجموعه‌ای از کلمات به نام Hits تبدیل می‌شود (رکوردهای Hits شامل اطلاعات اضافی در مورد اسناد و صفحات مثل رنگ، ظرفیت، قلم و غیره هستند). همچنین Indexer همه پیوندهای یک صفحه وب را تجزیه و اطلاعات مهم در مورد پیوندها را در واحد Anchors ذخیره می‌کند. Sorter تمام سندهای موجود در Barrels را بر اساس DocIDها مرتب و سپس عملیات نمایه معکوس را برای صفحات وب ایجاد می‌کند.

Barrels: این واحد پایگاه داده نمایه‌ساز محسوب می‌شود و نتایج حاصل از نمایه‌سازی را که به صورت رکوردهای Hits هستند، در خود ذخیره می‌کند. Anchors: این واحد تمام پیوندها و اطلاعات مربوط به پیوندهای صفحات وب را به تفکیک مشخص و ذخیره می‌کند و نشان می‌دهد که هر پیوند از کجا سرچشمه گرفته یا به کجا اشاره می‌کند.

Lexicon: نمایه‌ساز پس از اینکه صفحات وب را تجزیه کرد، تمام کلمات (شامل کلمات کلیدی) و اطلاعاتی اضافی هر سند را در واحد Lexicon قرار می‌دهد. همچنین در این واحد یک جدول درهم‌سازی³ نیز وجود دارد که هر رکورد آن به کلمات صفحات وب اشاره می‌کند.

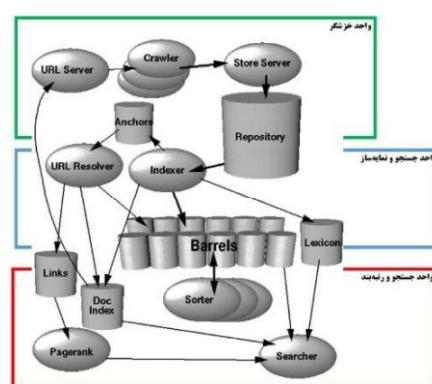
URL Resolver: این واحد فایل‌ها را از Anchors دریافت می‌کند و پس از تجزیه، آنها را به واحدهای دیگر ارسال می‌کند.

Links: Links پایگاه داده‌ای است که تمام پیوندهای ورودی و خروجی یک صفحه وب را در خود ذخیره می‌کند. این پیوندها برای رتبه‌بندی صفحه وب به واحد Pagerank فرستاده می‌شود (زیرا پیوندها بیشترین تأثیر را در مؤلفه رتبه‌بند جویش‌گرها برای رتبه‌بندی صفحات وب دارند).

² Module

³ Hash Table

جویش‌گر انجام و نتایج جستجو ارائه می‌شود، کاربران نتیجه کار بخش‌های متفاوت آن را می‌بینند. جویش‌گر از قبل پایگاه داده‌اش را آماده کرده است و این گونه نیست که درست در همان لحظه جستجو، تمام وب را بررسی کند. گوگل و هیچ جویش‌گر دیگری توانایی انجام این کار را ندارند. همه آن‌ها در زمان پاسخ‌گویی به جستجوهای کاربران، تنها در پایگاه داده‌ای که در اختیار دارند به جستجو می‌پردازنند و نه در کل اینترنت. جویش‌گر به کمک بخش‌های متفاوت خود، اطلاعات مورد نیاز را از قبل جمع‌آوری، تجزیه و تحلیل و سپس آن را در پایگاه داده‌اش ذخیره و به هنگام جستجوی کاربران، تنها پایگاه داده خودش را بررسی می‌کند. مؤلفه‌های مجزا و اصلی یک جویش‌گر به چندین بخش تقسیم می‌شود که عبارتند از [2]، [3]، [4]: ربات و عنکبوت¹، خزش‌گر، نمایه‌ساز، پایگاه داده و رتبه‌بند. به طور معمول هر جویش‌گر از معماری متفاوتی برای انجام جستجوها و پاسخ به درخواست‌های کاربران استفاده می‌کند؛ اما در این قسمت، یک معماری کلی و عمومی که مؤلفه‌ها و بخش‌های آن در بیشتر جویش‌گرها وجود دارد، بررسی می‌شود. شکل (۱) چنین معماری‌ای را نشان می‌دهد (جویش‌گر گوگل نیز از این معماری استفاده می‌کند) [5].



(شکل-۱): معماری عمومی جویش‌گرها [5]

مهم‌ترین بخش‌هایی که در هر یک از سه مؤلفه خزش‌گر، نمایه‌ساز و رتبه‌بند تعریف می‌شوند، در ادامه تشریح شده است.

Store Server و Repository دریافت فهرست URL‌ها آنها را واکشی و نتیجه کار را به مؤلفه Store Server ارسال می‌کند. Store Server پس از

¹ Spider

۲- تهدیدها و آسیب‌پذیری‌های مؤلفه‌های جویش‌گرهای بومی

در این بخش تهدیدها و آسیب‌پذیری‌های سه مؤلفه اصلی خزش‌گر، نمایه‌ساز و رتبه‌بند در جویش‌گرهای بومی بررسی شده است.

۲-۱- تهدیدها و آسیب‌پذیری‌های مؤلفه خزش‌گر

خزش‌گرهای وب کار اصلی کاوش داده‌ها و اطلاعات را بر عهده دارند و جویش‌گرها براساس داده‌های خزش‌شده به کاربران پاسخ می‌دهند. در ادامه برخی از انواع حملات، تهدیدها و آسیب‌پذیری‌هایی که به‌طور مستقیم یا غیر مستقیم روی خزش‌گرها تأثیر دارند، تشریح شده است [8].

۲-۱-۱- فاکتورهای طراحی خزش‌گر

پارامترهای طراحی و نیازمندی‌های مربوط به مؤلفه خزش‌گر، تأثیر زیادی در فرآیند پاسخ‌گویی به پرس‌وجوهای کاربران دارند. به‌طور مثال اگر جنبه‌های امنیتی برای مسائلی مانند پنهانی باند، ساختمان داده، کنترل کننده‌ها، سامانه پردازشی و منابع ذخیره‌سازی در خزش‌گرها لحاظ نشود، ممکن است، خزش‌گر در زمان مناسب وظایف خود را نتواند انجام دهد و عدم توجه به مسائلی از این قبیل، تهدیدی برای مؤلفه خزش‌گر محسوب می‌شوند [8، [9].

۲-۱-۲- حملات منبع خدمت

به‌طورمعمول خزش‌گرها برنامه‌هایی هستند که پیوسته صفحات وب را بررسی می‌کنند و به‌محض اینکه وب‌سایت یا صفحه جدیدی ایجاد شد یا صفحات تغییر کردند، تغییرات را با انجام عملیاتی خاص، در قسمت نمایه‌سازی اعمال می‌کنند؛ بنابراین اگر این برنامه‌هایی کاربردی برای مدت زمان کوتاهی غیر فعال شوند، پاسخ به پرس‌وجوهای کاربران بر اساس صفحات و اطلاعات قدیمی انجام خواهد شد؛ علاوه‌بر این ممکن است، طی این زمان صفحاتی آلوده ایجاد شوند؛ به‌همین دلیل لازم است با هر عاملی که باعث وقفه در کار خزش‌گر می‌شود، مقابله شود؛ چون مؤلفه‌های دیگر از خروجی خزش‌گر استفاده می‌کنند، وقفه در کار خزش‌گر باعث می‌شود مؤلفه‌های زیرین نیز تحت تأثیر قرار گیرند.

Doc index نگهداری می‌کند. علاوه‌بر Doc index واحد Barrels نیز ID صفحات را نگهداری می‌کند (زیرا هر صفحه وب با یک شماره ID شناسایی می‌شود).

Pagerank آن است. در Pagerank صفحات وب براساس پارامترهایی مثل پیوندها (پیوندهای ورودی و خروجی) رتبه‌بندی می‌شوند و خروجی این مرحله به سمت واحد Searcher فرستاده می‌شود.

Searcher یا واحد جستجو، مؤلفه نهایی است که از اطلاعات سایر واحدها (Lexicon) برای نمایش بهتر نتایج و پاسخ به درخواست‌های کاربران استفاده می‌کند.

با توجه به توضیحاتی که برای هر کدام از بخش‌های معماری جویش‌گرها ارائه شد، سه مؤلفه اصلی خزش‌گر (خزش‌گر از بخش‌های اصلی "سیاست‌های خزش، فاکتورهای طراحی، حملات تزیریک" که دسترس‌پذیری خزش‌گرها و پایگاه داده خزش‌گرها تشکیل شده است)، نمایه‌ساز (بخش‌های اصلی نمایه‌ساز " واحد جستجو، پایگاه داده نمایه‌ساز، سازوکارهای نمایه‌سازی، فاکتورهای طراحی" است) و رتبه‌بند (سیاست‌های رتبه‌بندی، سیاست دفاعی در برابر بهینه‌سازی منفی جویش‌گرها یا سئو منفی^۱ و وب‌گاه‌های آلوده^۲) تهدیدهایی دارند که در صورت نپرداختن به ضعف‌های امنیتی، عملکرد جویش‌گرها کاهش می‌یابد و درنهایت منجر به افشای اطلاعات غیر عمومی وب سایتها، افشای اطلاعات خصوصی کاربران، خزش و نمایه‌شدن صفحات وب سایتها آلوده، بدخواه^۳ و مخرب، ارائه نتایج غیر مرتبط و عدم محبوبیت جویش‌گر می‌شود. آمارها نشان می‌دهد که جویش‌گرهای بزرگی مثل گوگل در برابر برخی از تهدیدهای خزش‌گر و رتبه‌بند آسیب‌پذیر بوده‌اند [6، [7]. طبق بررسی‌های صورت‌گرفته جویش‌گرهای بومی مثل پارسی‌جو کمودهایی در زمینه سه مؤلفه اصلی دارند که تهدیدی برای آنها به شمار می‌رود؛ بنابراین استفاده از الزامات امنیتی ارائه شده در این مقاله به روند ارتقای جنبه‌های امنیتی جویش‌گرهای بومی کمک می‌کند.

¹ Negative Search Engine Optimization(SEO)

² Pollution Site

³ Malicious

دو فصل نامه علمی ترویجی منادی امنیت فضای تولید و تبادل اطلاعات (افتا)

رفتار ربات‌ها و بایگانی کردن صفحات وب‌گاه توسط جویش‌گر را طبق پروتکل‌های پذیرفته شده زیر می‌توان کنترل کرد.
این پروتکل‌ها عبارتند از [10]:

Robots.txt, XML Sitemap, Robots Meta Tag, Rel=Nofollow

پروتکل‌های Robots.txt و XML Sitemap برای کل وب‌گاه تعریف می‌شوند، به عبارتی سیاست‌های کلی وب‌گاه را تعریف می‌کنند. Robots Meta Tag برای هر صفحه وب به طور جداگانه لحاظ می‌شود و Rel=Nofollow به طور مجزا برای هر پیوند در نظر گرفته می‌شود. بیشتر جویش‌گرها این پروتکل‌ها را رعایت می‌کنند؛ اما هیچ‌کدام مجبور به رعایت آنها نیستند. در واقع این پروتکل‌ها مواردی را به جویش‌گرها توصیه می‌کنند و رعایت و عدم رعایت اینها به جویش‌گرها بستگی دارد؛ زیرا جویش‌گرها به خودی خود جسور و اگر مؤدب^۲ باشند (ربات‌ها، اسپایدرها و خوش‌گرها) برای حفظ حریم خصوصی وب‌گاه‌ها، این توصیه‌ها را رعایت می‌کنند. بهتر است که جویش‌گرها، توصیه تمامی وب‌گاه‌هایی را که از پروتکل‌های بالا استفاده کرده‌اند، رعایت کنند؛ زیرا مالکان وب‌گاه‌ها ممکن است اطلاعاتی را روی وب‌گاه قرار دهند که برای حفظ حریم خصوصی، تمایلی برای خرش موقت آن صفحات نداشته باشند. به عنوان مثال اگر نشانی یک وب‌گاه تغییر کرده است و مالک وب‌گاه تمایلی ندارد که جویش‌گرها صفحات قدیمی را خرش کنند، این تغییرات را در XML Sitemap و Robots.txt اعمال می‌کند. به همین دلیل جویش‌گرها برای اعمال تغییرات وب‌گاه‌ها و عدم خرش صفحات منع شده برای حفظ حریم خصوصی وب‌گاه‌ها، بهتر است که توصیه‌های پروتکل‌های بالا را رعایت کنند. اگر جویش‌گرها توصیه‌های پروتکل‌های بالا را رعایت نکنند، ممکن است مهاجمان از صفحاتی که به طور مؤدب خرش نشده‌اند، سوء استفاده کنند، به عنوان مثال دستیابی مهاجمان به صفحات پاک شده یک وب‌گاه که حاوی اطلاعات مهم است. به طور کلی اگر خوش‌گرها توصیه‌های وب‌گاه‌ها را رعایت نکنند یا اگر سیاست‌های خرش مناسب برای خرش صفحات وب نداشته باشند، انواع تهدیدها مثل تله‌اسپایدر و حلقه بی‌نهایت برای خوش‌گرها اتفاق می‌افتد [8].

برخی از سیاست‌های کلی و عمومی خوش‌گرها در مواجه با وب‌گاه‌ها عبارتند از [10، 11]: ۱- سیاست

حملات منبع خدمت^۱ یکی از تهدیدهایی هستند که منجر می‌شود جویش‌گرها نتوانند سرویس مناسب را در زمان تعیین شده ارائه کنند؛ زیرا زمانی که مؤلفه خوش‌گر تحت حمله منع خدمت است، ساختار کلی جویش‌گر تحت تأثیر قرار می‌گیرد و خدماتی که جویش‌گر ارائه می‌کند، براساس محتواها و داده‌هایی است که در قبل آنها را خوش کرده است؛ به عنوان مثال زمانی که یک کاربر سعی می‌کند جستجویی در مورد یک پایگاه داده که محتوای آن خیلی سریع به روزرسانی می‌شود، داشته باشد، جویش‌گر به استبانه محتواهایی را که از قبل پایگاه داده در خود موجود دارد، به کاربر نمایش می‌دهد.

۲-۱-۳- تزریق کد به ربات‌های خوش‌گر

امروزه نفوذ‌گرها برای رسیدن به اهداف خود از ابزارها و ایده‌هایی استفاده می‌کنند که با کمترین ایجاد ردپا به بالاترین نتیجه دست پیدا می‌کنند؛ به طوری که آمار شرکت‌های امنیتی نشان می‌دهد، آنها برای حملات تزریق SQL از ربات‌های جویش‌گرها استفاده می‌کنند؛ زیرا با این عمل ربات‌های جویش‌گرها، حملات تزریق SQL را روی پایگاه‌های هدف انجام می‌دهند. با آلوهشدن ربات‌های جویش‌گر ممکن است، یک پایگاه از طرف یک جویش‌گر معابر مثل گوگل مورد حمله قرار گیرد؛ در این شرایط اگر پایگاه مورد حمله، ربات‌های جویش‌گر را مسدود کند، آنگاه این عمل مانع نمایه‌سازی پایگاه خواهد شد و در غیر این صورت باید به ترافیک مخرب تولیدشده از سمت جویش‌گر اجازه فعالیت بدهد. در واقع هدف نفوذ‌گرها از آلوهشدن ربات‌های جویش‌گرها، پاک کردن ردپای خود (گمنامی هکرها) و انجام عملیات خرابکارانه به صورت موققیت‌آمیز روی وب‌سایتها است. به عنوان مثال مهاجم به جای اینکه خود پیوندی از حملات SQLI ایجاد و مابقی عملیات هک را به ربات‌های جویش‌گر واگذار می‌کند. بدین ترتیب ربات‌های جویش‌گرها با اجرای پیوند SQLI تولید شده توسط نفوذ‌گر، وب‌گاه site.com را هدف حمله قرار می‌دهند [7].

۲-۱-۴- نقض حریم خصوصی وب‌گاه‌ها و سیاست‌های خوش

² Politeness

¹ Denial-Of-Service (DOS)

طبق بررسی‌های صورت‌گرفته آپاچی سُلر در بیشتر جویش‌گرهای از جمله جویش‌گرهای بومی استفاده می‌شود و مهمترین آسیب‌پذیری آپاچی سُلر مربوط به حمله پیمایش فهرست^۱ است. سناریوی حملات پیمایش فهرست بدین صورت است که خدمت‌گزارهای وب نیز مانند ساختار فایلی که در سیستم‌عامل وجود دارد، دارای سلسله‌مراتب هستند، خدمت‌گزارهای وب به‌طورمعمول دسترسی به ریشه فهرست‌های خود را محدود می‌کنند؛ کاربران در این حالت به فهرست‌های زیرمجموعه می‌توانند دسترسی پیدا کنند؛ اما به مرحله بالاتر که ریشه است نمی‌توانند دسترسی پیدا کنند؛ همچنین نمی‌توانند به صورت موازی از پوشش‌های کنار دستی نیز استفاده کنند و تنها به محلی ارجاع داده می‌شوند که به آن دسترسی دارند. برخی اوقات به علل مختلف که بارزترین آن‌ها سهل‌انگاری مدیر خدمت‌گزار و ب است، دسترسی‌های مناسب برای پوشش ریشه در نظر گرفته نمی‌شود و همین امر می‌تواند باعث شود مهاجم به فهرست ریشه خدمت‌گزار و ب دسترسی پیدا کند. پیمایش فهرست سوءاستفاده از همین نقاط ضعفی است که در سطوح دسترسی یا عدم اعتبارستجی داده‌های ورودی کاربر به‌ویژه نام فایل‌ها در خدمت‌گزار و ب است که در نهایت می‌تواند باعث شود مهاجم با وارد کردن نویسه‌های مشخصی به فهرست ریشه^۲ دسترسی پیدا کند. هدف این‌گونه حملات دسترسی پیدا کردن به پوشش‌هایی است که نبایستی به صورت عمومی در دسترس قرار بگیرند، در این حمله در نهایت نفوذ‌گر می‌تواند فایل‌های موجود در قسمت محدود شده^۳ را دست کاری کند و نتیجه لازم را بگیرد.

۲-۲-۲-۲- فاکتورهای طراحی نمایه‌سازی و موازی‌سازی فرآیندها

یکی از مهم‌ترین مسائلی که باید در واحد نمایه‌ساز جویش‌گرهای امنیتی در نظر گرفت، فاکتورهای طراحی است؛ زیرا این فاکتورها عامل اصلی سرعت و عملکرد نمایه‌ساز در قبال پرس‌وجوهای کاربران است. برخی از مهم‌ترین فاکتورها و پارامترهای اصلی که در طراحی معماری نمایه‌ساز جویش‌گر استفاده می‌شود، عبارتند از: فاکتورهای ادغام، فاکتورهای ساختمان داده‌ای، اندازه نمایه‌سازی، روش‌های ذخیره‌سازی،

انتخاب^۴: این سیاست به‌طور کلی وضعیت دانلود صفحات وب را مشخص می‌کند. ۲- سیاست بازمشاهده^۵: این سیاست تغییرات صفحات وب را کنترل می‌کند. ۳- سیاست ادب^۶: این سیاست برای جلوگیری از بارگذاری دوباره صفحات، اعمال کردن تغییرات وب گاهها و حفظ حریم خصوصی آنها استفاده می‌شود. ۴- سیاست موازی‌سازی^۷: این سیاست نشان می‌دهد که خوش‌گرهای توزیع شده چگونه باید با یکدیگر هماهنگ شوند.

۲-۲-۳- تهدیدهای آسیب‌پذیری‌های مؤلفه نمایه‌ساز

مؤلفه نمایه‌ساز وظیفه جمع‌آوری، ذخیره‌سازی و تجزیه داده‌ها را برای بازیابی درست و آسان اطلاعات بر عهده دارد. بدون نمایه‌سازی جویش‌گرهای باید زمان زیادی را صرف جستجو در پایگاه داده خود کنند.

۲-۲-۴- تهدیدهای واحد جستجو وظیفه اصلی واحد جستجو پاسخ به درخواست‌ها و پرس‌وجوهای کاربران است و این واحد موظف است درخواست‌های جستجو را از کاربران دریافت و آنها را در قالبی مناسب جهت شروع فرآیند جستجو سازماندهی کند. جویش‌گرهای به‌طورمعمول برای واحد جستجو می‌توانند از هر بسترهای مثل آپاچی سُلر^۸، Compass، ElasticSearch، Ferret، SwiftType، DocFetcher و غیره (این سکوهای نرم‌افزاری روی نمایه‌ساز آپاچی لوسین پیاده‌سازی می‌شوند) استفاده کنند. آپاچی سُلر و ElasticSearch برای پرس‌وجوهای زیاد در مقیاس بزرگ عملکرد بسیار خوبی دارند؛ اما هر کدام از این دو مورد آسیب‌پذیری‌هایی دارند که در صورت سوءاستفاده از آنها مهاجم به راحتی به کل پایگاه داده از راه دور دسترسی می‌تواند داشته باشد. برخی از آسیب‌پذیری‌های واحد جستجو عبارتند از [۹]، [۱۰]، [۱۱]: Cross-site request forgery (CSRF)، Directory traversal (Dir.Trav)، Cross-site scripting (XSS)، Exec Code Inject

¹ Selection Policy

² Re-Visit Policy

³ Politeness Policy

⁴ Parallelization Policy

⁵ Apache Solr

⁶ Directory Traversal

⁷ Parent Directory

⁸ Restricted

دفاعی آنها به صورت دستی و خودکار^۱ در فرایند جستجو استفاده می‌شوند. سایر تهدیدهای سئوی منفی و الزامات امنیتی آن که گوگل نیز از آنها بهره می‌برد، در قسمت مؤلفه رتبه‌بند تشریح شده است.

۲-۲-۵- امنیت پایگاه داده‌ها

پایگاه داده یکی از واحدهای مهم برای مؤلفه نمایه‌ساز بهشمار می‌آید؛ زیرا مؤلفه نمایه‌ساز تمام اطلاعاتی را که از بخش خوش‌گر دریافت می‌کند، نمایه‌سازی و درنهایت خروجی حاصل از نمایه‌سازی را در پایگاه داده قرار می‌دهد. برخی از تهدیدهای امنیتی پایگاه داده‌ها که شامل جویش‌گرها نیز می‌شود، عبارتند از: سوءاستفاده از دسترسی‌های اضافی، سوءاستفاده از دسترسی‌های قانونی، ترفع حق دسترسی، آسیب‌پذیری‌های پلتفرم، آسیب‌پذیری‌های استنتاجی پایگاه داده، تزریق کد، فعلابودن ویژگی‌های غیرضروری DBMS، سرربیزی بافر، عدم رمزنگاری داده‌ها و نبود اطلاعات پشتیبان.

۲-۳- تهدیدها و آسیب‌پذیری‌های مؤلفه رتبه‌بند بهینه‌سازی جویش‌گرها یکی از مهم‌ترین روش‌هایی است که مالکان وب‌گاه‌ها برای افزایش رتبه و بهبود حضور وب‌گاه از آن استفاده می‌کنند. در صورت استفاده درست از سئو رتبه وب‌گاه‌ها در بخش رتبه‌دهی جویش‌گرها افزایش در غیر این صورت، سئو اثر معکوس روی رتبه‌دهی وب‌گاه‌ها دارد. حملات مربوط به سامانه رتبه‌بندی جویش‌گرها بیشتر با نام سئوی منفی شناخته می‌شوند و هدف این حملات تغییر سامانه رتبه‌دهی و رتبه‌بندی جویش‌گرها است. برخی از روش‌های سئوی منفی که در [۱۴]، [۱۵] به عنوان سئوی کلاه‌سیاه^۲ شناخته می‌شوند، در ادامه تشریح شده است.

۲-۳-۱- حمله ربودن پراکسی یا هک پراکسی

روش کار حمله ربودن پراکسی بدین صورت است که ابتدا مهاجم یک رونوشت از صفحات وب‌گاه هدف (وب‌گاه قربانی) را روی یک خدمت‌گزار پراکسی ایجاد و سپس با استفاده از روش‌های دیگر سئوی کلاه‌سیاه و پیوندهای کمکی، پیوندی

تحمل‌پذیری خطأ، سرعت جستجو و غیره [۹]. یکی از چالش‌های اساسی دیگری که جویش‌گرها با آن مواجه هستند، مدیریت فرآیندهای پردازش سریالی است. به عنوان مثال وقتی یک سند جدید به مجموعه اسناد اضافه می‌شود، نمایه‌ساز باید به روزرسانی شود؛ اما نمایه‌ساز در آن زمان نیاز دارد تا به پرس‌وجوهای کاربران نیز پاسخ دهد. بنابراین اینجا دو وظیفه رقابتی (وظیفه به روزرسانی و وظیفه پاسخ به پرس‌وجوها) اتفاق می‌افتد که نمایه‌ساز باید آنها را مدیریت کند. این چالش در ذخیره‌سازی‌ها و پردازش‌های توزیع شده بیشتر آشکار می‌شود.

۲-۲-۳- زمان‌بندی و دوره به روزرسانی نمایه‌ساز

پاسخ به درخواست‌های کاربران بر اساس نمایه‌سازی‌های به روز شده یک امر مهم در جستجو است؛ به طوری که باید همیشه سرویس‌ها بر اساس اطلاعات تازه و به روز ارائه شود. علاوه‌بر مسئله تازگی صفحات وب، به این نکته مهم باید توجه داشت که اگر یک صفحه وب در زمان T نمایه شده باشد و مهاجم در بازه زمانی T تا T+n صفحه را تغییر دهد و یک صفحه آلوده ایجاد کند و جویشگر تغییراتی را که در بازه زمانی T تا T+n اتفاق می‌افتد، ندیده باشد، آنگاه ممکن است کاربران به صفحه‌ای که مهاجم در بازه زمانی T تا T+n (صفحه آلوده) ایجاد کرده است، هدایت شوند. بنابراین اگر دوره به روزرسانی نمایه‌سازی صفحات زیاد باشد، آنگاه جویش‌گر نمی‌تواند صفحه آلوده و تغییر کرده را شناسایی کند.

۲-۲-۴- سئوی منفی

اگر چه سئوی منفی بیشتر مؤلفه رتبه‌بند جویش‌گرها را تهدید می‌کنند؛ اما یک جویش‌گر اگر عکس العمل خوبی در برابر سئوی منفی نداشته باشد، آنگاه مؤلفه نمایه‌ساز جویشگر، نیز تحت تأثیر این حمله قرار می‌گیرد. به عنوان مثال مهاجم با استفاده از سئوی منفی، وب‌گاه رقیب خود (قربانی) را هدف قرار می‌دهد؛ سپس جویش‌گر در این مرحله در صورتی که راه‌کار مناسبی در زمینه سئوی منفی نداشته باشد، ابتدا رتبه وب‌گاه قربانی را کاهش می‌دهد و به دلیل کاهش رتبه درنهایت برخی از صفحات وب‌گاه قربانی را ممکن است از سامانه نمایه‌سازی خود حذف کند، به همین دلیل جویش‌گرهای بزرگی مثل گوگل در زمینه مقابله با سئوی منفی خیلی خوب عمل کرده‌اند و امروزه سیاست‌های

^۱ Automatic

^۲ Black Hat SEO

اما جویش‌گرهای دیگر از جمله جویش‌گرهای بومی سازوکار خاصی در این زمینه ندارند.

۲-۳-۲- مزرعه پیوند

مزرعه پیوند^۳ یکی دیگر از روش‌هایی است که مالکان وب‌گاه برای دست‌کاری سامانه رتبه‌بندی جویش‌گرها استفاده می‌کنند. در واقع مزرعه پیوند به گروهی از وب‌گاه‌ها گفته می‌شود که بدون اینکه ارتباط مفهومی با هم داشته باشد، دست به مبالغه گسترده پیوند می‌زنند تا محبوبیت خود را افزایش دهند. در نگاه نخست مزرعه پیوند یک پیوند معمولی از یک وب‌گاه به نظر می‌رسد؛ در حالی که اینگونه نیست و در بیشتر موارد به طور تصادفی، ناخواسته یا هدف‌دار از وب‌گاه‌های دیگر سرچشمه می‌گیرد. در واقع اگر درجه یا چگالی پیوندهای ورودی یا خروجی یک وب‌گاه از یک مقدار آستانه بیشتر باشد یا پیوندهای ورودی از وب‌گاه‌های سرچشمه بگیرند که رتبه پایین و محتوای نامناسب دارند، (مثل Bad Links & Bad Neighbours, Link Spammering) سامانه رتبه‌بندی جویش‌گرها برای وب‌گاه مقصد تغییر می‌کند. امروزه جویش‌گرها براساس یکسری ویژگی‌های خاص، صفحات وبی را که با مزرعه پیوند مرتبط هستند، شناسایی می‌کنند و در نتایج جستجو نشان نمی‌دهند. البته در برخی از موارد ممکن است، فقط رتبه وب‌گاه کاهش یابد (برخی از استراتژی‌های سئوی موجود در جویش‌گرها، برای مقابله با مزرعه پیوند استفاده می‌شود).

۲-۳-۴- ریسندگی محتوا

ریسندگی محتوا^۴ یکی دیگر از روش‌هایی است که برای بهینه‌سازی جویش‌گرها استفاده و اما استفاده نامناسب از آن باعث تغییر رتبه وب‌گاه‌ها، در سامانه رتبه‌بندی جویش‌گرها می‌شود. ریسندگی محتوا فرایند بازنویسی یک محتوا (عبارت، صفحه وب و غیره) از روی یک محتوای اصلی است؛ به‌گونه‌ای که رونوشت ایجادشده باعث ایجاد محتوای تکراری نشود. در واقع هدف از ریسندگی محتوا ایجاد یک‌صد یا هزار محتوای ایجادشده جدید از روی محتوای اصلی، با نسخه‌های گوناگون و در زمان‌های متفاوت است؛ زیرا در صورت تکراری شدن محتوای صفحات وب به وسیله ریسندگی محتوا، حمله‌ای مانند ربدون پراکسی برای وب‌گاه اصلی اتفاق

از وب‌گاه‌های معتبر دیگر به صفحات موجود در خدمت‌گزار پراکسی ایجاد می‌کند و این عمل در درازمدت منجر به افزایش رتبه صفحات موجود در خدمت‌گزار پراکسی می‌شود. بنابراین جویش‌گرها وقتی دو وب‌گاه با محتوای یکسان را ملاقات می‌کنند، رتبه دو وب‌گاه را در سامانه رتبه‌بندی تغییر می‌دهند؛ زیرا جویش‌گرها ممکن است به‌دلیل تکراری بودن صفحات، صفحات و وب‌گاه معتبر را از سامانه شاخص‌دهی خود حذف یا رتبه پایین‌تری به صفحات وب‌گاه معتبر تخصیص دهند. اگر چنین اتفاقی برای وب‌گاه معتبر بیفتند، مهاجم به هدف خود که همان کاهش رتبه صفحات وب‌گاه معتبر بوده، رسیده است (به عنوان مثال نشانی وب‌گاه معتبر می‌تواند www.example.com و نشانی وب‌گاه جعلی www.example.net که روی خدمت‌گزار پراکسی قرار گرفته باشد). هدف از حمله ربدون پراکسی^۱ تغییر صفحات یک وب‌گاه معتبر به صفحات غیر معتبر است؛ زیرا با این عمل رتبه صفحات یک وب‌گاه در نتایج و مؤلفه رتبه‌بند و نمایه‌ساز جویش‌گر تغییر می‌کند. همچنین یک مهاجم ممکن است با ربدون پراکسی به یکسری اطلاعات مهم از وب‌گاه رقیب دسترسی پیدا کند، یا درنهایت درخواست کاربران وب‌گاه رقیب را به صفحات جعلی و آلوده (صفحات شامل بدافزار، تروجان و غیره) هدایت کند.

۲-۳-۵- چاشنی کلمه کلیدی

چاشنی کلمه کلیدی^۲ یک روش غیر مجاز در سئو است و باعث می‌شود که یک وب‌گاه به‌طور موقت یا دائمی از نتایج جویش‌گرها حذف شود. چاشنی کلمه کلیدی زمانی اتفاق می‌افتد که تعداد، چگالی و تنوع کلمات کلیدی موجود در متاتگ‌ها یا صفحات یک وب‌گاه بیش از حد زیاد باشد، از این طریق سعی می‌شود که رتبه وب‌گاه را به‌طور مصنوعی افزایش دهد. کلمات کلیدی که در چاشنی کلمه کلیدی HTML استفاده می‌شوند، بیشتر به‌طور مخفی در کدهایی مربوط به قلم، رنگ، عکس و غیره قرار می‌گیرند. چاشنی کلمه کلیدی به‌همراه حملات دیگر مثل ربدون پراکسی نیز استفاده می‌شود. بر اساس بررسی‌های صورت‌گرفته در منبع [۳۳] جویش‌گر گوگل هر وب‌گاهی را که از چاشنی کلمه کلیدی استفاده کند، شناسایی و رتبه آن را کاهش می‌دهد؛

¹ Proxy Hijacking (or Proxy Hacking)

² Keyword stuffing

وپرس‌ها و تروجان‌ها آلوده می‌شود، خزش‌گر جویش‌گر که به احتمال بالایی آن صفحات وب را کشف کرده، ممکن است آنها را در نتایج جستجو نشان دهد. در چنین مواردی یک جویش‌گر ایمن به جای اینکه مستقیم صفحات آلوده را به کاربران نشان دهد، آن را با پیامی مبنی بر اینکه "این سایت" یا صفحه وب ممکن است، برای سامانه شما خطناک باشد" به کاربران هشدار می‌دهد و کاربران با اختیار خود به آن صفحات رجوع می‌کنند.

۲-۳-۸- امنیت وب‌گاه

امروزه جویش‌گرها وب‌گاه‌هایی را که مورد خزش قرار می‌دهند از لحاظ امنیتی وارسی و نتیجه آن را در سامانه رتبه‌بندی اعمال می‌کنند. بنابراین اگر یک وب‌سایت کمینه امنیت لازم را در برابر انواع حملاتی مانند نشت اطلاعات، SQLI، XSS و غیره نداشته باشد، رتبه آن وب‌گاه در سامانه رتبه‌بندی جویش‌گرها تغییر خواهد کرد (امنیت وب‌گاه‌ها توسط ابزارهای خاص جویشگر ارزیابی می‌شود)، به‌طوری که جویش‌گر گوگل در سال ۲۰۱۴ اعلام کرد در زمینه امنیت وب‌سایتها، سیاست‌های جدیدی را در سامانه رتبه‌بندی خود اعمال می‌کند. بنابراین وجود سیاست‌های امنیتی برای ارزیابی سطح امنیت وب‌گاه‌ها برای اعتبار و محبوبیت جویش‌گرها یک امر الزامی است.

۳- الزامات و سیاست‌های امنیتی برای مؤلفه‌های اصلی جویش‌گرهای بومی

در این بخش تمام سیاست‌ها و الزامات امنیتی که جویش‌گرها باید برای سه مؤلفه اصلی خود یعنی خزش‌گر، نمایه‌ساز و رتبه‌بند در نظر بگیرند، تشریح شده است؛ زیرا در صورتی که جویش‌گرها سیاست‌های امنیتی کافی را در مؤلفه‌های اصلی خود نداشته باشند، ممکن است حملات مختلفی مثل سئوی منفی، حملات منع خدمت و حملات تزریق کد، جویش‌گرها را تهدید کند و این تهدیدها درنهایت باعث نمایش نتایج غیر مرتبط، Crash کردن جویش‌گرها و عدم محبوبیت آنها می‌شود.

۱- تدوین الزامات و سیاست‌های امنیتی خزش‌گر

جویش‌گرها به‌طور معمول بر اساس اطلاعاتی که خزش‌گرها

می‌افتد؛ درواقع اگر ریسندگی محتوا به‌طور پیوسته انجام شود، درنهایت منجر به حمله ریودن پراکسی خواهد شد.

۲-۳-۵- صفحه پنهان

صفحه پنهان^۱ روشی در بهینه‌سازی جویش‌گرها است؛ که در آن محتوای ارائه شده به اسپایدر یا ربات جویش‌گر، با چیزی که کاربر در مرورگر خود مشاهده می‌کند، متفاوت است. صفحه‌ای که به ربات جویش‌گر نشان داده می‌شود، ممکن است از لحاظ چگالی کلمات کلیدی، پیوندهای ورودی و خروجی، نوع محتوا و غیره مناسب باشد؛ اما صفحه‌ای که به کاربر نشان داده می‌شود، ممکن است یک صفحه هرز آگهی، غیر اخلاقی و بی‌ارزش باشد. بنابراین با حمله صفحه پنهان، صفحات هرزنامه در رتبه‌بند جویشگر دست‌خوش تغییر نمی‌شود و فقط صفحه نشان داده شده به ربات‌ها با افزایش یا کاهش رتبه مواجه می‌شود. درواقع وب‌گاه‌ها برای ارائه محتواهای متفاوت به کاربران و جویشگرها از نشانی IP یا header User-Agent HTTP درخواست‌ها استفاده می‌کنند.

۶-۳- صفحه درگاه

صفحات درگاه^۲ به‌طور معمول شامل صفحاتی هستند که هر کدام برای یک کلیدواژه یا عبارت خاص بهینه‌سازی شده است و هیچ‌گونه ارزش محتوایی ندارند. در بسیاری از موارد به‌خصوص برای بالا بردن رتبه صفحات یا عبارات خاص در جویش‌گرها طراحی شده‌اند و درنهایت کاربران را به یک سایت از قبل مشخص شده، هدایت می‌کنند. صفحات درگاه برای کاربران بسیار بد است؛ زیرا باعث می‌شود که کاربران به صفحات مشابه متعددی که صفحه مقصد آنها یکسان است، هدایت شوند. به این معنی که اگر کاربر روی هر پیوند موجود در صفحه درگاه کلیک کند، مقصد همه پیوندها یک صفحه از قبل تعیین شده است. این صفحات چه بر مبنای یک دامنه و چه چند دامنه باشد، تنها خستگی را برای کاربر به همراه می‌آورند.

۷- ۲-۳-۷- وب‌گاه آلوده^۳

موقعی که یک وب‌سایت توسط برنامه‌های مخرب مثل

¹ Cloacking Page

² Doorway pages

³ Pollution Web Site

وارسی پیوندها، استفاده از فهرست سفید است. در این روش ابتدا یک ساختار اعتبارسنجی قوی پیاده‌سازی می‌شود که ورودی‌ها و پیوندهای قانونی و معتبر را می‌پذیرد.

۳-۱-۲- الزامات امنیتی فاکتورهای طراحی

پارامترهای طراحی و نیازمندی‌های مربوط به مؤلفه خزش‌گر تأثیر زیادی در فرآیند پاسخ‌گویی به پرس‌وچوهای کاربران دارند و عدم توجه به مسائلی از این قبیل، تهدیدی برای مؤلفه خزش‌گر محسوب می‌شوند. برخی از فاکتورهای امنیتی مورد نیاز در مؤلفه خزش‌گر عبارتند از:

- حل کننده DNS: استفاده از حل کننده DNS آسنکرون چندنحوی

- خدمت‌گزارهایی با منابع مناسب: استفاده از لوح‌های سخت سریع مثل SSD، اسکارزی، ساس^۲ با حجم ذخیره‌سازی بالا
- پهنانی باند بالا و کافی برای خزش بهتر صفحات وب
- مدیر یا کنترل کننده‌ها: اعمال سیاست‌های ملاقات وеб‌گاه‌ها، حذف و بهروزرسانی سیاست‌های خزش
- الگوریتم‌های خزش: استفاده از الگوریتم‌های جستجوی عرضی

۳-۱-۳- مقابله با حملات منع خدمت

حملات منع خدمت/توزیع شده یکی از عواملی هستند که منجر می‌شود خزش‌گرهای سامانه جویش‌گر نتوانند سرویس مناسب را در زمان تعیین شده ارائه کند؛ بنابراین داشتن راهکارهای لازم برای رفع چنین حملاتی یکی از الزامات امنیتی اساسی برای جویش‌گر محسوب می‌شود. برای جلوگیری از حملات منع خدمت در سطح خزش‌گر نیاز است که از نرمافزارهای امنیتی، مانند دیوار آتش‌های معمولی یا Network Layer Firewall و Web Application Firewall (Firewall) استفاده شود. بهره‌گیری از دیوار آتش‌های مبتنی بر مدل سند شی^۳ توصیه می‌شود. این دیوار آتش‌ها علاوه بر کشف حملات جدید سربار کمتری را نیز به شبکه وارد می‌کند.

۳-۱-۴- کاربردهای توصیفی امنیتی خزش‌گرهای

برخی از کاربردهای امنیتی خزش‌گرهای خزش را در موارد زیر

جمع‌آوری می‌کنند، به درخواست‌های کاربران پاسخ می‌دهند. بنابراین مؤلفه خزش‌گر با خزش درست و بهینه وب، اطلاعات و داده‌های بهتری در اختیار سایر مؤلفه‌های جویش‌گر می‌تواند قرار دهد. در ادامه برخی از الزامات امنیتی کلی که باعث امن شدن خزش‌گرهای می‌شود، ارائه شده است.

۱-۳- الزامات امنیتی حملات تزریق کد

برای جلوگیری از حملات تزریق کد به خزش‌گرهای جویش‌گرها باید یکسری سیاست در الگوریتم‌های خزش اعمال کنند تا پیوندها قبل از خزش وارسی شوند. در واقع جویش‌گرها بهتر است، ابتدا پیوند اصلی را خزش و سپس ادامه پیوند را در صورتی خزش کنند که هیچ نویسه و کلمه کنترلی در آن وجود نداشته باشد. نحوه تزریق یک کد بدین صورت است که ابتدا مهاجم مانند حالت زیر، پیوندی را در وب‌گاه هدف قرار می‌دهد و خزش‌گر وقتی به چنین پیوندی برخورد کرد، سعی می‌کند محتوای داخل این پیوند را خزش کند؛ اما رجوع به پیوند زیر باعث اجرای یک دستور SQL روی پایگاه داده وب‌گاه www.example.com می‌شود (این یک نمونه ساده از حمله تزریق کد بود، حالتهای پیشرفت‌تری نیز از این حمله وجود دارد که مهاجمان از آنها می‌توانند استفاده کنند).

http://www.example.com/index.php?id=number+union+n+delete+from+table_name...

برای وارسی پیوندها از دو رویکرد فهرست سیاه (یا سازوکار رهایی^۱) و فهرست سفید می‌توان استفاده کرد. در روش فهرست سیاه ابتدا فهرستی از نویسه‌هایی که باید ترجمه شوند، ایجاد می‌شود و در فهرست سیاه قرار می‌گیرند؛ سپس ورودی‌های خزش شده با این فهرست سیاه مقایسه می‌شوند. حال اگر در داده‌های ورودی نویسه‌ای وجود داشته باشد که در فهرست سیاه هم موجود باشد، آنگاه آن نویسه به یک نویسه بی‌معنی تبدیل می‌شود یا ممکن است آن نویسه حذف و ادامه پیوند خزش شود. استفاده از این روش خطراتی دارد؛ چون ممکن است، بعضی از نویسه‌های کنترلی در فهرست سیاه موجود نباشند؛ همچنین این روش بهدلیل مقایسه همه ورودی‌ها با فهرست سیاه پیچیدگی زمانی اش زیاد است. یکی دیگر از سازوکارهای دفاعی قابل قبول برای مسئله

¹ Escaping

² Serial Attached SCSI (SAS)

³ Document Object Model (DOM)

می‌کند. جهت اجتناب بارگیری بیش از یکبار صفحات، سامانه‌های خزش‌گر، نیازمند سیاستی هستند که نشانی‌های جدید کشفشده طی خزش را انتساب دهند. در این صورت، اگر یک نشانی یکسان توسط دو فرآیند خزش پیدا شود، مشخص خواهد شد و از دوباره کاری جلوگیری می‌شود؛ زیرا با وجود سیاست‌های موازی جنبه امنیتی دسترسی‌پذیری خدمت‌گزارهای خزش‌گر نیز بیشتر خواهد شد [11].

سیاست ادب: اگر یک خزش‌گر چندین درخواست در ثانیه انجام دهد و یا فایل‌های بزرگی را بارگیری کند، سرویس‌دهنده^۱ زمان زیادی را صرف پاسخ‌گویی به درخواست‌های خزش‌گر خواهد کرد و درنهایت این کار باعث می‌شود که کاربری سرویس‌دهنده کمی تغییر کند. بنابراین استفاده از خزش‌گرها برای کاربردهایی مفید است؛ ولی هزینه‌هایی هم به دنبال دارد. اگر چه وب‌گاه‌ها برای ارائه و نمایش خدمات خود به مشتریان، باید به درخواست‌های خزش‌گر جویش‌گرها پاسخ دهند، به‌طوری‌که بیشتر مالکان وب‌گاه‌ها از روش‌هایی استفاده می‌کنند تا خزش‌گرهای جویش‌گرها و بگاه آنها را بیشتر ملاقات کنند؛ اما به‌طورکلی خزش‌گرها (خزش‌گرهای وب که شامل خزش‌گر جویش‌گرها است) باعث اعمال برخی از هزینه‌ها مثل اتلاف منابع شبکه، سربار پردازشی و غیره به خدمات‌دهنده‌ها می‌شوند [11]. یک راه حل برای رفع بخشی از هزینه‌های تحمیلی به خدمات‌دهنده‌ها، استفاده از پروتکل ممانت از ربات‌ها است، این پروتکل با نام Robots.txt شناخته می‌شود. در همین اواخر، جویش‌گرهای تجاری بزرگ مانند Ask، Google، MSN و Yahoo! Search Jeeves نام Crawl-Delay در فایل Robots.txt می‌توانند استفاده کنند که این پارامتر تعداد ثانیه‌های تأخیر بین درخواست‌های خزش‌گرها از یک سایت را نشان می‌دهد؛ زیرا اگر زمان بین درخواست‌ها خیلی کم باشد، آنگاه خزش‌گر یک جویش‌گر ممکن است یک حمله منع خدمت را روی یک وب‌گاه انجام دهد [11].

۲-۳-۲- تدوین الزامات و سیاست‌های امنیتی نمایه‌ساز

نمایه‌سازی نامن اطلاعات و داده‌ها، تهدیدی بزرگی برای

می‌توان خلاصه و از نتایج آنها در سطوح نمایه‌ساز و رتبه‌بند استفاده کرد.

- کشف صفحات و وب‌گاه‌های آلوده در هنگام خزش وب؛
- کشف و مقابله با سئوی منفی در هنگام خزش وب؛
- ارائه راهکارهای امنیتی به وب‌گاه‌هایی که در سطوح پایین امنیت قرار دارند.

۳-۱-۵- سیاست‌های امنیتی خزش

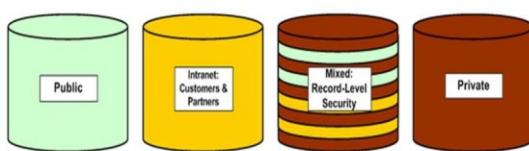
همان‌طورکه در قبل ذکر شد، خزش‌گر از چند سیاست اصلی برای خزش محتوای وب‌گاه‌ها استفاده می‌کند که در ادامه نحوه بهره‌گیری مناسب این سیاست‌ها به‌طور کامل تشریح شده‌اند.

سیاست انتخاب: اولویت‌بندی و انتخاب صفحات مهم یک ضرورت محسوب می‌شود. باستی صفحات وب بر اساس یک معیار اولویت‌بندی شوند و خزش‌گر وب بر این اساس حرکت کند تا مؤثرترین جمع‌آوری در کمترین زمان صورت گیرد. استراتژی‌های متفاوتی برای انتخاب محتواها توسط خزش‌گرها وجود دارد که در بخش استراتژی‌های خزش خزش‌گرها ذکر خواهد شد [11].

سیاست بازمشاهده: از دید یک جویش‌گر عدم شناسایی رویدادها که منجر به داشتن رونوشت منسخه شده از منابع می‌شود، هزینه دارد. تازگی و سن به‌طورمعمول جزء پراستفاده‌ترین توابع هزینه برای مقابله با رونوشت‌های قدیمی هستند. دو معیار تازگی و سن بدین منظور استفاده می‌شود تا بتوان تعیین کرد چه صفحاتی نیاز به مشاهده‌ی مجدد دارند. بنابراین زمان بازدید دوباره صفحات برای خزش‌گرهای یک جویشگر باید به‌گونه‌ای باشد که پاسخ به پرس‌وجوهای کاربران بر مبنای اطلاعات به‌روزشده انجام شود. سه رویدادی که ممکن است رخ دهد و باعث تغییر صفحات وب شود عبارتند از [11]: ۱- ایجاد یک صفحه جدید که نیازمند اضافه‌شدن پیوندی جدید است. ۲- تغییر و به‌روزرسانی صفحات که ممکن است، تغییر جزئی یا بزرگ باشد. ۳- حذف یک صفحه که باعث می‌شود، خروجی خزش‌گر پیوندی را نگه‌داری کند که دیگر وجود ندارد.

سیاست موازی: خزش‌گر موازی خزش‌گری است که چندین فرآیند را به‌صورت موازی اجرا می‌کند. هدف آن بیشینه‌کردن نرخ بارگیری است؛ در حالی‌که سربار موازی‌سازی را کمینه و از بارگیری صفحات تکراری اجتناب

^۱ Server



(شکل-۲): امنیت در سطح سند و جمع آوری [28]

صحت داده: به طور معمول در پایگاه داده‌های مبتنی بر SQL برای کنترل صحت داده‌ها بیشتر از محدودیت‌ها و رهانای^۱ استفاده می‌شود. همچنین در پایگاه داده‌های بزرگ از الگوریتم‌های داده‌کاوی، برای استخراج دانش از پایگاه داده می‌توان استفاده کرد و بر اساس دانش حاصله میزان صحت داده‌های ذخیره شده را تعیین کرد.

کنترل دسترسی: کنترل دسترسی باید در سلسله سطوح مختلف پایگاه داده لحاظ شود تا کاربران و واحدهای سیستمی بر اساس وضعیت و اهمیت خود بتوانند به قسمت‌های مورد نیاز از پایگاه داده دسترسی داشته باشند. برای این منظور باید ابتدا کمینه امتیازهای اصلی که برای کاربران نیاز است، در سامانه تعریف و پیاده‌سازی شود؛ سپس در مراحل بعدی برای هر کاربر یا واحد بر اساس اعتبار و جایگاهی که در طول زمان کسب می‌کند، تفویض اختیار و امتیاز صورت گیرد. سیاست‌های کنترل دسترسی که برای پایگاه داده نمایه‌ساز جویش‌گرها تعریف می‌شود، بیشتر به واحدهای سیستمی داخل جویش‌گر برمی‌گردد. برای داشتن یک پایگاه داده امن باید موارد زیر به طور پیوسته کنترل تا تهدیدهای مربوطه در بالاترین سطح، خنثی و امنیت در سلسله‌مراتب پایگاه داده حفظ شود:

- کنترل دسترسی‌های قانونی برای جلوگیری از سوء استفاده؛
- اعمال سیاست‌های امنیتی برای جلوگیری از ترفعی حق دسترسی؛
- شناسایی آسیب‌پذیری‌های پلتفرم و وصله کردن آنها؛
- کنترل استنتاج‌های منطقی؛
- محدود کردن پارامترهای ورودی برای جلوگیری از تزریق کد
- پشتیبان‌گیری از اطلاعات؛
- غیر فعال کردن ویژگی‌های غیر ضروری پایگاه داده‌ها

نوع پایگاه داده: به طور معمول جویش‌گرها به دلیل

محرمانگی داده‌ها و اطلاعات یک وب‌گاه است. بنابراین داشتن راه کارهای لازم در زمینه نمایه‌سازی امن محتوا، برای هر وب‌گاهی الزامی است. از آنجا که جویش‌گرها نیز اطلاعات وب‌گاهها را نمایه می‌کنند باید نمایه‌سازی خیلی امن تری داشته باشند. در بیشتر سرویس‌ها، نمایه‌سازی باید به صورت سلسله‌مراتبی انجام شود؛ یعنی ابتدا اطلاعات ورودی تجزیه و تحلیل شده، سپس با استفاده از الگوریتم‌های فشرده‌سازی و رمزگاری، اطلاعات فشرده و رمز شوند و درنهایت در فهرست مربوطه ذخیره شدن. در ادامه سایر واحدهای استفاده شده برای نمایه‌سازی تشریح شده‌اند.

۱-۳-۲-۱- ذخیره‌سازی داده‌ها در پایگاه داده

نمایه‌سازی باید به صورت سلسله‌مراتبی انجام شود؛ یعنی ابتدا اطلاعات ورودی تجزیه و تحلیل، سپس با استفاده از الگوریتم‌های فشرده‌سازی و رمزگاری، اطلاعات فشرده و رمز و درنهایت در فهرست مربوطه ذخیره می‌شوند. در فرایند نمایه‌سازی اگر عملیات رمزگاری اطلاعات مهم انجام نشود، مهاجم با استفاده از حملات مربوط به فهرست نامن، فهرست مربوطه را کشف و به راحتی اطلاعات غیر مجاز را از فهرست کشف شده، بازیابی و مشاهده می‌کند.

ذخیره‌سازی امن داده‌ها در پایگاه داده مستلزم توجه به جنبه‌های مختلفی چون محرمانگی و صحت داده‌ها، کنترل دسترسی و نوع پایگاه داده است، که در ادامه به هر یک از این جنبه‌ها اشاره می‌کنیم.

محرمانگی: روش‌هایی برای حفاظت از محرمانگی و پنهان‌سازی داده‌های ذخیره شده در پایگاه داده‌ها وجود دارد که اغتشاش داده‌ها، رمزگاری داده‌ها، تسهیم داده‌ها و بازیابی محرمانه اطلاعات، محدود کردن پرس‌وجوها از مهم‌ترین آنها هستند. در فرایند نمایه‌سازی داده‌ها، برای افزایش امنیت باید از گسسته‌سازی اجزای ذخیره‌سازی و پیاده‌سازی سازوکارهای کنترل دسترسی استفاده کرد تا نشت اطلاعاتی صورت نگیرد؛ زیرا با گسسته‌سازی اجزا اگر دسترسی غیر مجازی به داده‌ها صورت گیرد، فقط بخشی از داده‌ها آشکار خواهد شد (همچنین در گسسته‌سازی اجزا سیاست‌های امنیتی متفاوتی را روی هر جزء ذخیره‌سازی، می‌توان پیاده‌سازی کرد) [28]. [16]. شکل (۲) نحوه سطح‌بندی، گسسته‌سازی و واحد بندی یا پیمانه‌ای کردن منابع ذخیره‌سازی را نشان می‌دهد و هر کدام از منابع برای داده خاصی استفاده می‌شود.

¹ Trigger

اعطای کمینه مجوز (Least Privilege)، صدور مجوز برای استفاده از داده‌ها و اطلاعات ذخیره‌شده برای افزایش پشتیان‌گیری از داده‌ها و اطلاعات ذخیره‌شده برای افزایش دسترس‌پذیری.^{۱۴} ۱۵- استفاده از ساختارهای امنیتی مثل هانی پات‌ها^۳ برای سابقه‌گیری^۴ و به دست آوردن اطلاعات بیشتر از مهاجمان و رائمه سناریوهای از پیش تعیین شده به آنها.^{۱۶} ۱۷- کنترل آسیب‌پذیری‌های پلتفرم و غیرفعال کردن خدمات اضافی.^{۱۷} ۱۸- کنترل ورودی‌های نامعتبر توسط توابع امنیتی کتابخانه‌ای، برای جلوگیری از حملات تزریق کد.^{۱۸} ۱۹- استفاده از پروتکل‌های امنیتی (SSL, TLS, IPsec) برای برقراری ارتباط امن بین خدمت‌گزارهای خوش‌گر و پایگاه داده.^{۱۹} ۲۰- محدودساختن هر واحد یا کاربر به داده هایی که مجاز به دسترسی یا تغییر آنها است (کنترل دسترسی).^{۲۰} ۲۱- اینمنی ساختار پایگاه داده در مقابل انتشار تغییرات ناخواسته (جامعیت منطقی پایگاه).^{۲۱} ۲۲- اینمنی داده های موجود در پایگاه داده نسبت به مخاطرات فیزیکی (جامعیت فیزیکی پایگاه).^{۲۲} ۲۳- دسترسی به پایگاه داده در همه شرایط (دسترس‌پذیری).^{۲۳} ۲۴- بررسی هویت کاربران در ارتباط با ردگیری نظارتی و اجازه دسترسی به داده‌های خاص (تصدیق اصالت و هویت شناسی کاربران).^{۲۴} ۲۵- توجه به مسائل سازگاری اطلاعات ذخیره‌شده در پایگاه داده (استفاده از محیط‌های ابری و مجازی‌سازی باعث افزایش ناسازگاری پایگاه داده می‌شود که این مسائل باید توسط تیم امنیت مدیریت شود).

۳-۲-۲- موازی‌سازی فرآیندها

جویش‌گرها به دلیل اینکه از معماری توزیع شده بهره می‌برند، بیشتر با این چالش مواجه هستند؛ زیرا اگر موازی‌سازی فرایندها به درستی کنترل نشود، زمانی که کاربر یک جستجو انجام می‌دهد، انتظار دارد در کمترین زمان ممکن پاسخ مناسب را دریافت کند و این زمان با موازی‌سازی فرایندها در جستجوهای پایگاه داده جلوگیری در بهینه‌ترین حالت قابل دستیابی است. بنابراین باید به مسائل موازی‌سازی و همزمان‌سازی توجه ویژه‌ای داشت. به طور معمول با استفاده از فریمورک‌هایی مثل هدوپ^۵ مسائل مربوط به مدیریت محتوا

³ Honeypot

⁴ Log

⁵ Hadoop

اینکه روزانه با داده‌های زیادی (جریانی از داده‌ها) مواجه می‌شوند نیاز دارند از پایگاه داده‌هایی استفاده کنند که عملیات ذخیره‌سازی داده‌ها در آنها سریع‌تر انجام شود و تکرار داده و همزمان‌سازی در آنها نیز به راحتی صورت گیرد. پایگاه داده‌های NoSQL برای این منظور مناسب هستند و برای ذخیره‌سازی داده‌های زیاد عملکرد خوبی دارند. در این پایگاه داده‌ها به دلیل اینکه لایه دسترسی به داده‌ها از لحاظ فیزیکی یا نرم‌افزاری کنترل می‌شود و تعداد سطوح دسترسی نسبت به پایگاه داده‌های SQL بیشتر است، سطح امنیتی بالاتری دارند و برای سرویس‌های جویش‌گرها که با داده‌ها و اطلاعات زیادی سروکار دارند، توصیه می‌شود [17].

به طور کلی سایر الزامات امنیتی که در پایگاه داده نمایه‌ساز و خوش‌گر می‌تواند استفاده شود، عبارتند از:
 ۱- محدود کردن کدهای پویای پایگاه داده‌ای.^۲ ۲- گسترش‌سازی و دانه‌بندی کردن منابع ذخیره‌سازی برای کاهش نشت اطلاعات در صورت نفوذ به سیستم.^۳ ۳- رمزگذاری فایل‌های داده برای برقراری محرمانگی داده‌ها.^۴ ۴- رمزگذاری و احرار هوتیت خدمت‌گزار و مشتری.^۵ ۵- رمزگذاری و احرار هویت بین خوشه‌ای.^۶ ۶- بهره‌گیری از روش‌های جلوگیری از تزریق کد WAVES, DBC-Checker, WebSSARI, Java Static مثل SQL DOM, Tainting, SecuriFly ذخیره‌سازی داده‌های موجود در هر عنصر پایگاه داده (صحت داده).^۷ ۷- دقت در (نظرات‌پذیری).^۸ ۸- ردگیری عامل انجام عملیات در پایگاه داده (نظرات‌پذیری).^۹ ۹- کنترل کردن اسکریپتها و ورودی‌های اجرایی مخرب و مقایسه آنها با فهرست‌های کنترلی از پیش تعیین شده.^{۱۰} ۱۰- ارائه ساختارها و واحدهای امنیتی مثل دیوار آتش به صورت سلسله‌مراتبی برای جلوگیری از ارتباط مستقیم به پایگاه داده‌ها.^{۱۱} ۱۱- اعتبارسنجی اندازه، قالب، بازه و نوع داده‌های ورودی.^{۱۲} ۱۲- مدیریت قوی کلیدهای پایگاه داده‌ای و بهره‌گیری از سامانه‌های مدیریت کلیدی که از واسط پایگاه داده و نوع کلیدهای آن، پشتیبانی کند. امروزه از واحدهای سخت‌افزاری امن مثل HSM^۲ برای این منظور استفاده می‌شود.^{۱۳} ۱۳- اعمال روش‌های کنترل حملات تزریق کد: پرس‌وجوهای پارامتر بندی شده (Parameter Query)، پردازه‌های ذخیره‌شده (Store Procedure)، استفاده از دستور Escaping، استفاده از فهرست سفید (White List).

¹ Auditability

² Hardware Security Module (HSM)

داده‌ای که برای نمایه‌سازی اطلاعات در جویش‌گرها می‌توان استفاده کرد، نمایه معکوس است.

۶-۳-۲-۶- دوره به روزرسانی نمایه‌ساز

رسیدن به یک مصالحه برای زمان دوره به روزرسانی نمایه‌ساز به طوری که تازگی اطلاعات همواره در حد مطلوب باشد، برای هر جویش‌گری یک الزام اساسی است. همچنین دوره به روزرسانی باید به گونه‌ای تنظیم شود که صفحات وب‌گاه‌های آلوده نیز در کمترین زمان ممکن کشف شوند. در بیشتر جویش‌گرهای بزرگ دوره‌های به روزرسانی نمایه‌ساز ثانیه‌ای یا ساعتی است.

۶-۳-۲-۷- سئوی منفی

در جویش‌گرهای بزرگی مثل گوگل و بینگ سیاست‌های تعریف شده است که اگر وب‌گاه‌ها مسائل مربوط به سئوی کلاه‌سیاه را در وب‌گاه‌ها لحاظ کنند، آنگاه رتبه آنها کاهش یافته و یا ممکن است از فهرست نمایه‌سازی حذف شوند. بنابراین اگر جویش‌گری در برابر سئوی منفی این نباشد، آنگاه احتمال حذف وب‌گاه‌های درست از فهرست نمایه‌سازی وجود دارد. بهمین دلیل جویش‌گرها باید سیاست‌های دفاعی خود را همیشه به صورت دستی یا خودکار در فرایند جستجو پیاده‌سازی کنند تا هیچ وب‌گاه درست و سالمی با مشکل مواجهه نشود.

۶-۳-۲-۸- سیاست‌های مقابله با صفحات و وب‌گاه‌های آلوده

به طور کلی در فرآیند نمایه‌سازی ممکن است، صفحاتی نمایه شوند که به هیچ‌وجه ارزش محتوایی برای کاربران نداشته باشند و رجوع به این صفحات نارضایتی کاربران را بهمراه دارد. مؤلفه نمایه‌ساز در صورتی که قابلیت کشف صفحات آلوده و مشکوک را داشته باشد، در زمان نمایه‌سازی با همکاری مؤلفه رتبه‌بند از دو رویکرد زیر می‌تواند استفاده کند.

- اگر جویش‌گر، یک صفحه وب را آلوده تشخیص داد، آن صفحه را می‌تواند نمایه‌سازی نکند و به طور کامل آن را کنار بگذارد. با این کار کاربران به هیچ‌وجه از طریق جویش‌گر به صفحه آلوده هدایت نمی‌شوند؛ اما مشکلی که در این رویکرد وجود دارد، مسئله تشخیص اشتباه است؛ زیرا اگر یک صفحه سالم به اشتباه، آلوده تشخیص

و پردازش‌های موازی روی معماری‌های توزیع شده به خوبی انجام می‌شود و در بیشتر زمان‌ها چالش‌های موازی‌سازی قابل کنترل می‌باشند.

۶-۳-۲-۳- فاکتورهای طراحی نمایه‌ساز پیاده‌سازی الزامات زیر در هنگام طراحی نمایه‌ساز جویش‌گرها باعث امن شدن معماری آن خواهد شد:

- قابلیت اطمینان نرم‌افزاری؛
- افزونگی محتوا؛
- توزیع شدگی واحدهای نمایه‌سازی برای افزایش دسترسی‌پذیری؛
- رفع خطاهای مربوط به نمایه‌سازی در انزوا؛
- افزایش تحمل پذیری خط؛
- خدمت‌گزارهای مناسب جهت کاهش هزینه‌های محاسباتی؛
- سامانه‌های پشتیبان؛
- مراکز داده مناسب

۶-۳-۲-۴- واحد جستجو

دلایل اصلی که باعث بروز آسیب‌پذیری در سکوهای واحدهای جستجو می‌شود، ضعف‌های نرم‌افزاری است که در هنگام توسعه به وجود آمداند. توصیه‌ای که برای این منظور می‌توان بیان کرد، اطمینان از به روز بودن بسته‌های نرم‌افزاری و آزمایش پیوسته آنها است تا در صورت وجود آسیب‌پذیری، سریعاً وصله امنیتی آن نصب شود. همچنین فیلتر کردن تمامی ورودی کاربران (به خصوص اجازه ندهید که متأ نویسه‌ها، داخل ورودی‌های کاربران قابل درج باشد). برای مقابله با کدهای اجرایی در واحد جستجو، داده‌های غیر قابل اعتماد از دستورها و پرس‌وجوها را می‌توان جدا کرد. به این نکته نیز باید توجه کرد که سیاست‌های دفاعی اپلیکیشن‌های تحت وب، قابل استفاده در واحد جستجو است.

۶-۳-۲-۵- ساختمان داده نمایه‌ساز

یکی از مسائل مهمی که در نمایه‌سازی جویش‌گر اهمیت دارد، نحوه انتخاب ساختمان داده نمایه‌ساز است. انواع مختلفی از ساختارها و ساختمان داده‌ها وجود دارد که برای نمایه‌سازی می‌توان استفاده کرد. مناسب‌ترین ساختمان

همکاری یکدیگر به بهترین نحو ممکن با هرزنامه‌ها می‌توانند مقابله کنند.

۳-۳-۲- سیاست‌های مقابله با ریسندگی محتوا و ربودن پراکسی

رویکردی که برای این منظور می‌توان استفاده کرد بدین صورت است، اگر یک صفحه در زمان T ایجاد شده و همان صفحه در زمان $T+n$ بازنویسی شده است، جویش‌گر باید زمان ایجاد صفحات را در نظر بگیرد تا پی‌برد که مسئله تکراری بودن و بازنویسی محتوا مربوط به کدام صفحه است. اگر زمان ایجاد صفحات به عنوان یک پارامتر در رتبه‌بندی صفحات لحاظ شود، آنگاه مسئله تکراری بودن صفحات وب و محتواها که به وسیله ریسندگی محتوا یا ربودن پراکسی ایجاد می‌شود، قابل حل خواهد بود. همچنین برای جلوگیری از هک پراکسی جویش‌گرها باید اتصالاتی را که از خدمت گزارهای پراکسی به وب‌گاه‌ها وارد می‌شود، در فرایند رتبه‌بندی صفحات لحاظ کنند و از تأثیر زیاد آنها در رتبه‌بندی جلوگیری شود؛ زیرا اگر پیوندهایی که از خدمت گزارهای پراکسی به وب‌گاه‌ها داده می‌شوند، به عنوان پیوند معمولی در نظر گرفته شود، آنگاه یک وب‌گاه یا صفحه هرزنامه با ربودن پراکسی به بالاترین رتبه در نتایج جویش‌گرها می‌رسد.

۳-۳-۳- سیاست‌های مقابله با صفحات پنهان

جویش‌گرها با سیاست و رویکرد تغییر نام یا تغییر نشانی IP اسپایدرها، می‌توانند وب‌گاه‌هایی را که محتوای متفاوتی ارائه می‌کنند، تشخیص دهند؛ سیاست تغییر نشانی با خوش‌گرهای توزیع شده که در نقاط جغرافیایی متفاوتی قرار گرفته‌اند محقق می‌شود؛ زیرا با این کار وب‌گاه‌ها جویش‌گر را به عنوان یک کاربر معمولی شناسایی می‌کنند. در این لحظه جویش‌گر چون از قبل محتوای متفاوتی را از همان وب‌گاه مشاهده کرده به راحتی پی‌می‌برد که صفحه مورد نظر یک صفحه پنهان است.

۳-۳-۴- سیاست‌های مقابله با مزرعه پیوند

- محدود کردن چگالی پیوندها برای رتبه‌بندی صفحات وب
- در نظر گرفتن مبدأ و مقصد پیوندها و شناسایی همسایه‌های

داده شود، آنگاه احتمال نمایه‌سازی آن صفحه سالم به طور تقریبی صفر است. بنابراین برای جلوگیری از این اتفاق رویکرد دوم پیشنهاد می‌شود.

- اگر جویش‌گر، صفحه وب را آلوده تشخیص داد می‌تواند با همکاری مؤلفه رتبه‌بند، آن صفحه را با میزان رتبه کمتری نمایه‌سازی کند؛ زیرا با این عمل کاربران با احتمال کمتری صفحات مشکوک و آلوده را مشاهده می‌کنند و اگر کاربران به پیوند صفحات مشکوک و آلوده رسیدند، جویش‌گر باید با صدور یک پیام هشدار، هدف چنین صفحاتی را برای کاربران آشکار کند.

۳-۳- الزامات و سیاست‌های امنیتی مؤلفه رتبه‌بند

در صورتی که مؤلفه رتبه‌بند یک جویش‌گر ضعیف و آسیب‌پذیر باشد، تهدیدی بزرگی برای وب‌گاه‌های سالم و درستکار به شمار می‌رود؛ زیرا چنین وب‌گاه‌هایی در زیر صفحات هرزنامه و مخرب دفن می‌شوند و کاربران به سختی قادر به مشاهده آنها هستند؛ بنابراین در این بخش سیاست‌ها و الزامات امنیتی کلی که باید در مؤلفه رتبه‌بند پیاده‌سازی شود، ارائه شده است.

۳-۳-۱- سیاست‌های مربوط به صفحات هرزنامه

هر صفحه وبی که از طریق روش‌های غیر مجاز مثل ریسندگی محتوا، مزارع پیوند، چاشنی کلمه کلیدی و غیره افزایش رتبه پیدا کند و در نتایج جستجو ظاهر شود، به عنوان صفحه هرزنامه شناخته می‌شود. مبارزه با صفحات هرزنامه در جویش‌گرها باید جزء اولویت‌های اصلی باشد؛ زیرا آنچه که باعث محبوبیت یک جویش‌گر می‌شود ارائه نتایج مرتبط و فریبنخوردن کاربران است. بنابراین می‌توان از الگوریتم‌های اکتشافی^۱ و ژنتیک^۲ که سرعت بالایی در پردازش دارند در تجزیه و تحلیل کدهای تحت وب برای کشف صفحات هرزنامه استفاده کرد. به طور معمول مقابله با هرزنامه‌ها در سطح رتبه‌بند صورت می‌گیرد؛ ولی منطقی‌ترین حالت پیاده‌سازی سیاست‌های امنیتی آن در سطوح مختلف (خرش‌گر، نمایه‌ساز و رتبه‌بند) جویش‌گر است به طوری که بخش نمایه‌ساز و رتبه‌بند مناسب‌ترین سطوح برای مقابله با هرزنامه‌ها است. در واقع این دو سطح با

¹Heuristic

²Genetic

شود، آنگاه شانس نمایه‌سازی آن به طور تقریبی صفر می‌شود به همین دلیل وجود سامانه دفاعی برای مقابله با صفحات آلوده در سطح رتبه‌بند، نیز پیشنهاد می‌شود. البته پیشنهاد نهایی یک حالت ترکیبی است که در قسمت سیاست‌های امنیتی نمایه‌ساز ارائه شد و با آن رویکرد، مقابله با صفحات آلوده بهترین عملکرد را خواهد داشت. به یقین مسئله رتبه‌بندی صفحات وب چیزی نیست که با درنظر گرفتن چند تهدید و اعمال راه‌کارهای امنیتی برای آنها بهترین رتبه‌بندی را بتوان داشت. زیرا مسئله رتبه‌بندی و الگوریتم‌های آن باید متناسب با زمان تغییر کند تا در برابر انواع تهدیدهای جدید این شود و در نهایت مناسب‌ترین صفحات در نتایج جستجو نشان داده شود [20], [19], [18].

۴- امنیت در معماری جویش‌گرهای

بومی

در بخش‌های قبل تهدیدهای و آسیب‌پذیری‌های مؤلفه‌های اصلی جویش‌گر مشخص و در همان راستا سیاست‌ها و الزامات امنیتی تعیین شدن؛ به طوری که با امن‌شدن مؤلفه‌ها، معماری کلی جویش‌گر نیز امن خواهد شد. به همین دلیل در این بخش امنیت معماری جویش‌گرها بر اساس مؤلفه‌ها تشریح شده است. امنیت در معماری جویش‌گرها باید در سه مرحله در نظر گرفته شود که این سه مرحله شامل مرحله واکشی، مرحله نمایه‌سازی و رتبه‌بندی در محیط خدمت‌گزارهای داخلی جویش‌گر و مرحله جستجو است. شکل (۳) معماری کلی و عمومی جویش‌گر را نشان می‌دهد. همان‌طور که در شکل (۳) مشاهده می‌شود، وقتی کاربران از طریق مرحله جستجو، پرس‌وجوهی خود را ارسال می‌کنند، قبل از اینکه به مرحله نمایه‌سازی برسد، باید توسط "مرحله پیش پردازش پرس‌وجوهی کاربران" وارسی و سیپس به سطح نمایه‌سازی ارسال تا مناسب‌ترین پاسخ به کاربران ارائه شود. همچنین در مرحله واکشی استناد نیز چنین ساختاری وجود دارد که قبل از اینکه استناد وب‌گاه‌ها به مرحله نمایه‌سازی برسد، با "مرحله پیش‌پردازش استناد" تجزیه و تحلیل می‌شود تا اصطلاحات و اطلاعات اشتباهی نمایه‌سازی نشوند؛ زیرا اگر چنین اطلاعات اشتباهی به مرحله نمایه‌سازی برسد، تطبیق دادن پرس‌وجوهای کاربران با آنها سخت‌تر می‌شود [23], [18].

- جریمه‌نکردن وب‌گاه‌های معتبر به دلیل پیونددادن از وب‌گاه‌های هرزنامه؛ زیرا مهاجم ممکن است به دلایلی از ربدون پراکسی استفاده و از وب‌گاه جعلی موجود در پراکسی که محتوای نامطلوب دارد، پیوندی به وب‌گاه اصلی و معتبر ایجاد کند. بنابراین اگر جویشگر در برابر چنین تهدیدی، سیاست و ساختار دفاعی لازم را نداشته باشد، آنگاه رتبه وب‌گاه اصلی به دلیل داشتن پیوند ورودی از همسایه بد تنزل خواهد کرد.

۵- سیاست‌های مقابله با چاشنی کلمه کلیدی

- اعمال چگالی ۷ الی ۱۰ درصدی از کلمات کلیدی برای

محاسبه رتبه هر صفحه وب [14]

- کشف کلمات کلیدی مخفی در کدهای وب و نادیده‌گرفتن آنها در رتبه‌بندی صفحات وب (در الگوریتم رتبه‌بندی جریمه‌هایی برای چنین کارهایی می‌توان اعمال کرد)

۶- سیاست‌های مقابله با صفحات درگاه

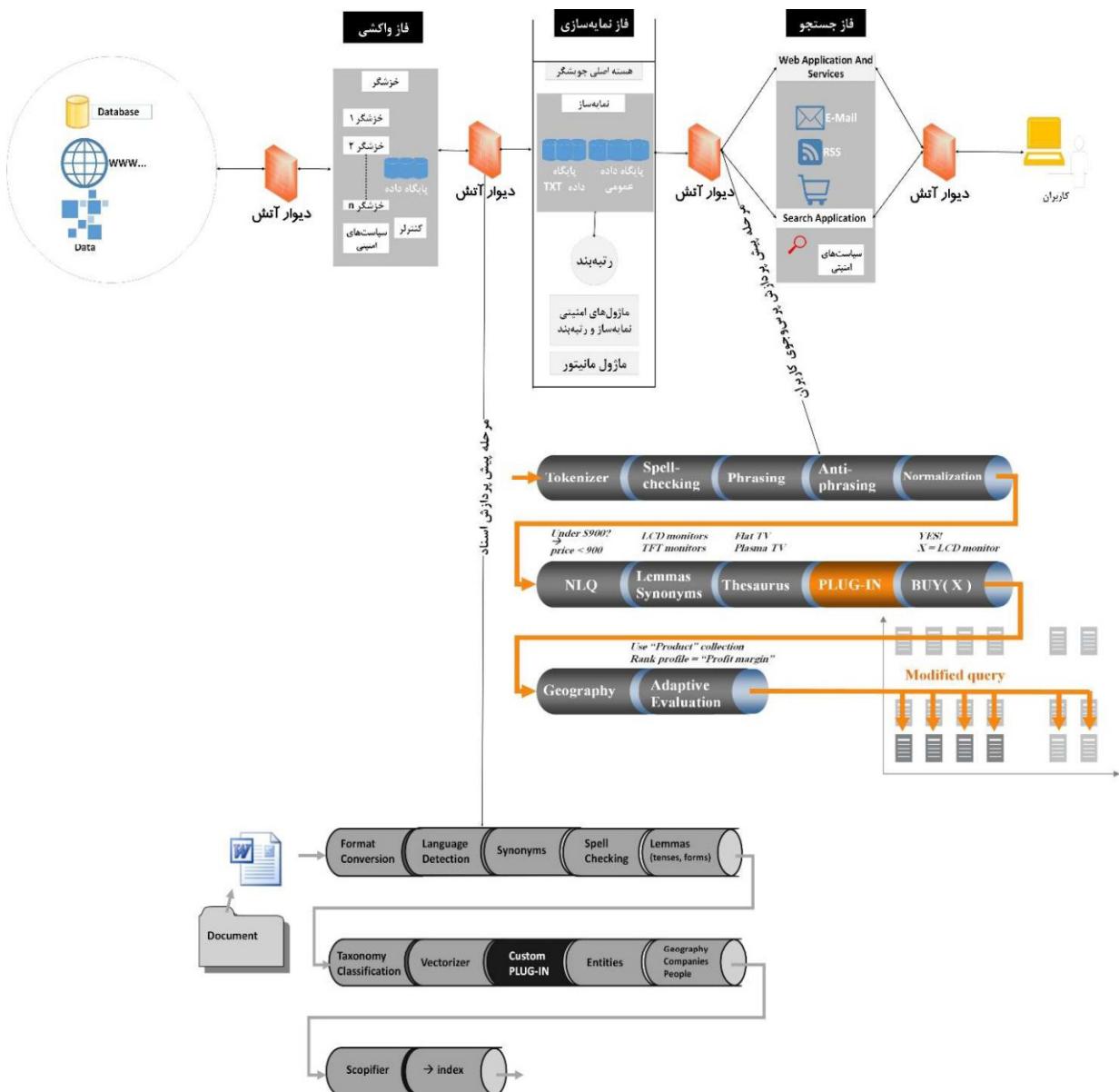
جویش‌گرها با تحلیل و پالایش کردن صفحات وب میزان چگالی بهینگی عبارات را می‌توانند محاسبه کنند و صفحاتی که روی یک عبارت و کلمه خاصی چگال‌تر هستند به عنوان صفحه درگاه شناخته شوند؛ بنابراین مهم‌ترین پارامتری که در شناخت چنین صفحاتی تأثیر دارد، فراوانی عبارات و کلمات خاص است. همچنین برخی از صفحات درگاه ممکن است، دارای اسکریپت‌های باشند که به محض اینکه کاربر به چنین صفحاتی رجوع می‌کند، آن اسکریپت اجرا شود. در واقع مبارزه با این صفحات باید با رویکردهای متفاوتی صورت گیرد که تمام حالت‌های مخرب توسط جویش‌گر شناسایی شود (هر سیاست امنیتی که برای صفحات هرزنامه استفاده می‌شود، قابل پیاده‌سازی برای این صفحات نیز است).

۷- سیاست‌های مقابله با صفحات و وب‌گاه‌های

آلوده

اگر چه مراحل مناسب برای کشف و مقابله با صفحات آلوده در زمان فعالیت خزش‌گر و نمایه‌ساز است؛ اما پیاده‌سازی سامانه دفاعی در مرحله پیش‌پردازش استناد نیز یک امتیاز می‌تواند محسوب شود؛ زیرا با این کار اگر یک وب‌گاه سالم به دلایل احتمالی و False Positive الگوریتم‌ها و سامانه‌های دفاعی در سطوح خزش‌گر و نمایه‌ساز، به اشتباه آلوده تشخیص داده

دو فصل نامه علمی ترویجی منادی امنیت فضای تولید و تبادل اطلاعات (افتا)



(شکل-۳): مراحل عمومی جویش‌گرهای بومی

دیوار آتش‌های معمولی یا هوشمند است؛ زیرا بیشتر محصولات امنیتی مثل دیوار آتش‌های لایه سه، چهار و هفت شبکه، سیاست‌های امنیتی کنترل پورت، تجزیه درخواست‌ها، شناسایی داده‌های مخرب، کنترل حملات منع خدمت، کنترل ترافیک‌های عبوری و غیره را انجام می‌دهند. شکل (۴) امنیت در مرحله واکشی را مشخص می‌کند؛ زیرا در این مرحله باید مسائل مربوط به سیاست‌های خوش، میازره با سئوی کلاه‌سیاه و وب‌گاه آلوده، امنیت پایگاه داده‌ها، فاکتورهای طراحی در نظر گرفته شود؛ زیرا پیاده‌سازی ساختارهای امنیتی در این مرحله باعث می‌شود که اثرات خراب‌کارانه آن به مراحل و سطوح دیگر

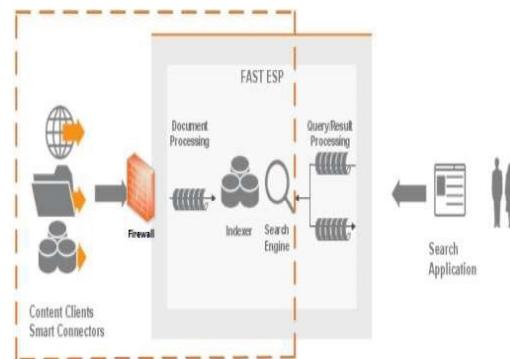
شکل‌های (۴) تا (۶) امنیت سلسله‌مراتبی را در مرحله واکشی، نمایه‌سازی و جستجو در این شکل‌ها با خط‌چین مشخص شده‌اند. مهم‌ترین مسئله امنیتی که در سه شکل مورد نظر، یک تهدید بزرگ برای جویش‌گر به‌شمار می‌رود، دسترسی به پایگاه داده‌های نمایه‌ساز است [۲۳]، [۲۴].

اگر ارتباطات از سمت مرحله واکشی و مرحله جستجو کنترل و مدیریت نشود، مهاجم از طریق این دو درگاه به پایگاه داده نمایه‌ساز دسترسی می‌تواند داشته باشد. مهم‌ترین واحد امنیتی مناسب برای این درگاه‌ها استفاده از

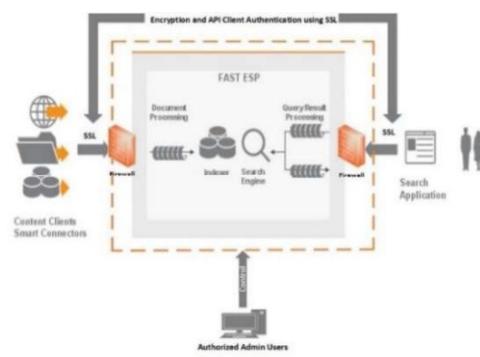
ساختارهای امنیتی آن همواره اهمیت دارد. مسائل امنیتی که باید در این قسمت به آن توجه داشت، شامل امنیت پایگاه داده، امنیت واحد جستجو، امنیت واحد تجزیه‌کننده، فاکتورهای طراحی و سیاست‌های مقابله با وب‌گاه‌های آلوده و سئوی منفی است. علاوه‌بر این مرحله نمایه‌سازی باید توسط واحدهایی مثل دیوار آتش از سمت مرحله واکشی و مرحله جستجو حفاظت تا میزان دسترسی‌های عمومی محدود شود. مهم‌ترین مسئله امنیتی که باید در مرحله نمایه‌سازی در نظر گرفت، امن‌کردن پایگاه داده نمایه‌ساز است؛ زیرا با دسترسی به پایگاه داده نمایه‌ساز امنیت جویش‌گر به خطر می‌افتد. بنابراین یکی از واحدهای اصلی امنیتی که باید برای این امر در نظر گرفت، استفاده از دیوار آتش است. شکل (۴) امنیت جویش‌گر را در مرحله اپلیکیشن جستجو نشان می‌دهد و کاربران نخستین سطحی را که در جویش‌گر می‌بینند، واسطه کاربری جستجو^۱ است. مسائل امنیتی که در مرحله جستجو باید به آن توجه شود، شامل سیاست‌های مقابله با حملات تحت وب (مثل SQLI، XSS، Buffer OverFlow، DOS و غیره است (شایان ذکر است که در این مقاله بررسی امنیت در مرحله جستجو یا اپلیکیشن تحت وب مد نظر نبوده و بیشترین تمرکز روی سه مؤلفه اصلی یعنی خزش‌گر، نمایه‌ساز و رتبه‌بند بوده است).

جدول (۱) انواع زیرواحدهای مؤلفه‌های اصلی جویش‌گرها را از ابعاد مختلف امنیتی بررسی می‌کند، در این جدول سعی شده است به گونه‌ای تکمیل شود که مؤلفه‌های هر جویش‌گری از جمله پارسی‌جو و یوز را شامل شود. این اطلاعات بر اساس استاندارد ITU-T X.805 که نگاهی کلی به ابعاد امنیتی اصلی دارد تکمیل شده است. طبق این جدول تمام زیرواحدها باید جنبه‌های امنیتی دسترس‌پذیری، کنترل دسترسی، تصدیق اصالت و امنیت ارتباط را فراهم کنند. دسترس‌پذیری باید در تمام چرخه حیات جویش‌گر همواره برقرار باشد تا جویش‌گر در هر شرایط و زمانی پاسخ‌گوی کاربران باشد.

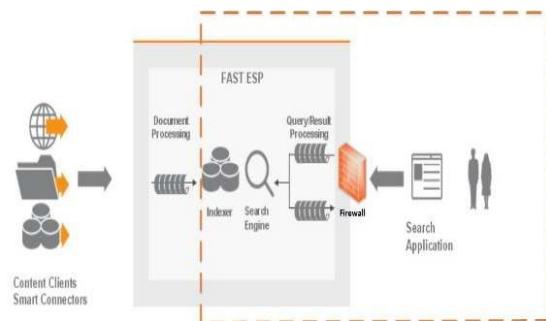
محرمانگی داده‌ها باید برای پایگاه داده خزش‌گر و نمایه‌ساز همواره برقرار باشد؛ زیرا اگر نشت اطلاعاتی صورت گیرد آنگاه مهاجمان نمی‌توانند به داده متن آشکار دسترسی داشته باشند، بنابراین وجود سازوکارهای رمزگاری برای



(شکل-۴): امنیت در مرحله واکشی [۲۴]



(شکل-۵): امنیت در مرحله نمایه‌سازی خدمت‌گزارهای داخلی [۲۴]



(شکل-۶): امنیت در مرحله اپلیکیشن جستجو [۲۴]

وارد نشود؛ زیرا با بهره‌گیری از ساختارهای امنیتی در سطوح و مراحل مختلف خطرات احتمالی آنها در همان سطوح خنثی می‌شود. در مرحله واکشی علاوه‌بر دیوار آتش از انواع سیاست‌های مقابله با سئوی منفی، صفحات و وب سایتهاي آلوده، سیاست‌های خوش برای کنترل تازگی اطلاعات و حذف صفحات آلوده، وارسی پیوندهای عبوری و غیره را می‌توان انجام داد.

شکل (۵) وضعیت ساختارهای امنیتی را در مرحله يا سطح نمایه‌سازی و رتبه‌بندی نشان می‌دهد؛ زیرا از آنجا که مرحله نمایه‌سازی مهم‌ترین قسمت و هسته اصلی یک جویش‌گر محسوب می‌شود، توجه به سیاست‌ها و

¹ Search Interface

دو فصل نامه علمی ترویجی منادی امنیت فضای تولید و تبادل اطلاعات (افتا)

اطلاعات دریافتی از کاربران به پرس‌وجوها پاسخ دهد، آنگاه باید مرحله گمنامی کاربران را در حین جستجو در نظر بگیرد تا ردپایی از کاربران روی وب سایتها ثبت نشود. همچنین الگوریتم‌های خزش باید حریم خصوصی وب‌گاهها را نیز رعایت کنند تا اطلاعات خصوصی که مالکان وب‌گاه مایل به نشان‌دادن آنها نیستند، افشا نشود (این مسئله بیشتر به سیاست‌های امنیتی الگوریتم‌های خزش جویش‌گر بستگی دارد).

جنبه امنیتی انکارناپذیری باید توسط الگوریتم‌های خزش حفظ شود؛ زیرا ممکن است که یک خزش‌گر به وب‌گاهی رجوع کند و به‌دلیل یک آسیب‌پذیری در خزش‌گر، حمله تزیق کد را روی وب‌گاه مربوطه انجام دهد. در این زمان مالک وب‌گاه بر اساس ردپای جویش‌گر باید به شناسایی آن باشد تا جویش‌گر نتواند عمل انجام‌شده را انکار کند.

برقراری محترمانگی همواره الزامی است. اگر چه خود الگوریتم‌های خزش‌گر، نمایه‌ساز و رتبه‌بند باید محترمانه باشند، اما ممکن است که یک جویش‌گر از الگوریتم‌های متن باز استفاده کند و در عمل محترمانگی برای آنها معنا نداشته باشد؛ ولی اگر الگوریتم‌ها بومی‌سازی شده باشد، آنگاه محترمانگی برای آنها حائز اهمیت خواهد بود.

صحت داده‌های ذخیره‌شده در پایگاه داده همواره مهم است؛ زیرا اگر اطلاعات متناقضی در پایگاه داده وجود داشته باشد، آنگاه پاسخ به پرس‌وجوها کاربران بر اساس اطلاعات اشتباہ خواهد بود؛ به‌طوری‌که اگر جویش‌گر از پایگاه داده‌های توزیع‌شده استفاده کند، وجود صحت داده اهمیت بیشتری برای پایگاه داده خواهد داشت. مسئله حریم خصوصی بیشتر در خدمات تحت وب اهمیت دارد و مؤلفه‌های اصلی جویش‌گر ارتباط کمتری با آن دارند؛ اما زیرا واحد جستجو و الگوریتم خزش باید به جنبه حریم خصوصی نیز توجه کنند؛ زیرا واحد جستجو اگر با

(جدول-۱): بررسی مؤلفه‌های جویش‌گرهای بومی از ابعاد امنیتی

رتبه‌بند	نمایه‌ساز												خزش‌گر				ابعاد امنیتی
	بازگردانی‌های ریزه‌بندی	بازگردانی‌های پیوسته	بازگردانی‌های معمولی	بازگردانی‌های انتخابی	بازگردانی‌های نمایه‌ساز	بازگردانی‌های احتمالی	کنندۀ	تحثیه	محاسبه	نمایه‌سازی							
دسترسی پذیری	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
محترمانگی	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓	✓
حریم خصوصی	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
کنترل دسترسی	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
صحت داده‌ها	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓	✓
تصدیق اصلت	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
امنیت ارتباط	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
انکارناپذیری	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗

۵- نتیجه‌گیری

امنیتی برای هر کدام از مؤلفه‌ها به‌طور مجزا ارائه شد و نشان داده شد که با امن‌شدن این سه مؤلفه، معماری کلی جویش‌گر امن خواهد شد. بخش پایانی مقاله به معماری امن جویش‌گر اشاره شد، در معماری پیشنهادی امنیت معماری

در این مقاله انواع حملات و آسیب‌پذیری‌هایی که هر کدام از مؤلفه‌های خزش‌گر، نمایه‌ساز و رتبه‌بند، یک جویش‌گر را افتاده تهدید می‌کنند، مورد بررسی قرار گرفت؛ سپس مناسب با منوچی آسیب‌پذیری‌های شناسایی‌شده، الزامات و سیاست‌های

علمی ترویجی

گوگل باید قابلیتی داشته باشد که کاربران بتوانند صفحات وب هرزنامه را به آنها گزارش دهند و با بررسی و تحلیل دقیق‌تر آن صفحات، تصمیم نهایی را برای حذف چنین صفحاتی از نتایج جستجو اتخاذ کنند. همچنین جویش‌گرهای بومی باید گزینه‌ای داشته باشد که به صورت برشخ^۳ میزان رتبه و درصد هرزنامه وب‌گاهها را نشان دهند.

۶- جویش‌گرهای بومی باید از زیرساخت‌های امن و قابل اطمینان برای سه واحد خوش‌گر، نمایه‌ساز و رتبه‌بند استفاده کنند تا بستر لازم برای جریان‌های داده‌ای برشخ زیاد فراهم شود.

۷- جویش‌گرهای بومی در استراتژی‌های انتخاب محتوا ضعیف هستند؛ بنابراین برای هر محتوایی و پاسخ به هر تقاضایی باید از استراتژی و خوش‌گر مربوط به همان محتوا استفاده شود. به عنوان مثال برای محتواهای علمی، عمومی، خاص و غیره رویکرد مجازی باید وجود داشته باشند.

به طور کلی به دلیل این که جویش‌گرهای از جمله جویش‌گرهای بومی از مؤلفه‌های زیادی برای ارائه سرویس استفاده می‌کنند و هر کدام از مؤلفه‌ها خود از چندین زیر مؤلفه (الجزایر کوچک‌تر) تشکیل شده است، امنیت در تمام زیر مؤلفه‌ها منجر به برقراری امنیت معماري جویش‌گر می‌شود و از دید کاربران یک جویش‌گر امن، جویش‌گری است که ۱- نتایج درست، به روز شده و مرتبط ارائه کند ۲- از ارائه وب‌گاه‌های آلوده و هرزنامه در نتایج جستجو جلوگیری کند ۳- هشدارها و پیغام‌های امنیتی را به کاربران قبل از مشاهده وب‌گاه‌های آلوده ارائه کند. بنابراین در این مقاله موضوع‌ها و محورهایی مورد بررسی قرار گرفت که پیاده‌سازی و اعمال آنها در جویش‌گرهای جویش‌گرها به همراه دارد.

۶- مراجع

- [1] P. Hunt and S. Hansman, "A Taxonomy of Network and Computer Attack," *Computers and Security*, pp.31-43, November 2003.
- [2] T. Y and G. A, "Web Search Engines: Practice and Experience," *Computing Handbook*, 2013.

² Online

جویش‌گر در سه مرحله اصلی به نام مرحله جستجو، مرحله نمایه‌سازی و رتبه‌بندی و مرحله واکشی بررسی و ارتباط ابعاد مختلف امنیتی با زیر واحدهای مؤلفه‌های اصلی جویش‌گر، ارائه شد. همچنین در ادامه پیاده‌سازی آنها منجر به بهبود عملکرد، افزایش محبوبیت، نتایج بهتر و ارتقاء امنیت (وب سایتها و کاربران) را برای جویش‌گرهای بومی به همراه دارد.

۱- پیاده‌سازی ابزارهای مدیریت وب‌گاهها مثل Google Webmaster Tools در جویش‌گرهای بومی؛ زیرا چنین ابزاری به مالکان وب‌گاه این اختیار را می‌دهد که وب‌گاه خود را متناسب با الگوریتم‌های رتبه‌بندی جویش‌گرهای بومی بهینه کنند و این ابزار برای مالکان وب‌گاه‌های تجاری بسیار جذاب خواهد بود؛ زیرا جویش‌گرهای بزرگی مثل گوگل سیاست‌های کلی پاداش و سیاست‌های جریمه الگوریتم‌های رتبه‌بندی خود را ارائه می‌کنند تا مالکان وب‌گاهها راحت‌تر در راستای آن سیاست‌ها بتوانند اقدام کنند.

۲- بهره‌گیری از پارامترهای رتبه‌بندی بیشتر برای رتبه‌دهی صفحات و وب‌گاه‌ها. زیرا طبق سند [۹] تعداد پارامترهایی که در جویش‌گرهای بومی استفاده می‌شود، نسبت به جویش‌گرهای جهانی کمتر است.

۳- ارائه ساختارها و سیاست‌های دفاعی مناسب برای مقابله با سئوی کلاه‌سیاه. زیرا طبق بررسی‌های صورت گرفته و بر اساس سند [۹]، سیاست مطلوبی که از سئوی کلاه‌سیاه جلوگیری کند، در ساختار مؤلفه‌های اصلی (مثل خوش‌گر، نمایه‌ساز و رتبه‌بند) جویش‌گرهای بومی وجود ندارد.

۴- جویش‌گرهای بومی در مواجه با صفحات سمی^۱ و آلوده، در مقایسه با جویش‌گرهای جهانی عملکرد و سیاست مناسبی ندارند و سیاست و ساختار امنیتی برای پالایش و تحلیل کدهای مخرب وب در آنها دیده نمی‌شود. همچنین ارائه هشدارهای امنیتی به کاربران در هنگام ملاقات صفحات آلوده یکی از الزامات اساسی برای جویش‌گرهای بومی به شمار می‌رود که باید اقدامات لازم برای پیاده‌سازی این مورد صورت گیرد.

۵- جویش‌گرهای بومی مانند جویش‌گرهای بزرگی مثل

¹ Poisoning

- using a Game Theoretic Approach," in Professional Communication Conference (IPCC), IEEE International, 2015.
- [16] E. Shmueli, R. Waisenberg, Y. Elovici and E. Gudes, "Designing secure indexes for encrypted databases," Data and Applications Security XIX. Springer Berlin Heidelberg, pp. 54-68, 2005 .
- [17] P. Noiumkar و T. Chomsiri, "A Comparison the Level of Security on Top 5 Open Source NoSQL Databases," Information and Communication Technology in the Faculty of Informatics of the Mahasarakham University, Thailand, 2014.
- [18] P. Niesten, "Legal status of search engine result manipulation," Tilburg Institute for Law, Technology, and Society (TILT), Netherlands, 2012.
- [19] "doorway-page" [Online]. Available: www.mozilla.org/ugc/protect-your-website-from-becoming-a-doorway-page-google-patent-style/, 2015.
- [20] "search_engine_security" [Online]. Available: www.pcworld.com/article/232224/search_engine_security.html.
- [21] Halfond William G., Jeremy Viegas, and Alessandro Orso. "A classification of SQL-injection attacks and countermeasures." Proceedings of the IEEE International Symposium on Secure Software Engineering. Vol. 1. IEEE, 2006.
- [22] Aryal, D and Shakya, A. "A Taxonomy of SQL Injection Defense Techniques", Sweden (2011).
- [۲۳] حسن کوشکی، شقایق نادری "امنیت در مؤلفه‌ها و معماری جویش‌گرها" کنفرانس بین المللی مهندسی کامپیوتر و فناوری اطلاعات، تهران، دانشگاه تهران، ۱۳۹۵/۰۳/۱۲.
- [23] H. Kooshkaki, S. Naderi, "Security in Components and Architecture Search Engines." International Conference on Computer Engineering and Information Technology, Tehran, Tehran University, 2016.
- [24] "FAST ESP Services - Search Technologies" [Online]. Available: www.searchtechnologies.com
- [25] "searchengineland" [Online]. Available: www.searchengineland.com.
- [26] "Search Engine Security" [Online]. Available: www.blog.sucuri.net
- [3] W. Croft, D. Metzler and T. Strohman, Search engines: Information retrieval in practice, United States: Addison-Wesley, Pearson Education, Inc,2015.
- [4] T. Y, G. A, "Web Search Engines: Practice and Experience," Computing Handbook, 2013.
- [5] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine", Computer networks, pp.3825-3833, 2012.
- [6] D. Thies, "Google Proxy Hacking: How A Third Party Can Remove Your Site from Google SERPs," Vancouver, Canada, 2007.
- [7] "google-bots-doing sql-injection-attacks," [Online]. Available: www.blog.sucuri.net/2013/11/google-bots-doing sql-injection-attacks.html.
- [8] O. Christopher and M. Najork, "Web crawling," Foundations and Trends in Information Retrieval, pp. 175-246, 2010.
- [۹] زارع بیدکی، علی محمد، طراحی و پیاده سازی سامانه‌های خزش و رتبه‌بندی اسناد فارسی وب و پیاده‌سازی یک جویشگر فارسی، ۸۹۳۲۲۴۱۲ پژوهشگاه فضای مجازی، پژوهشکده فناوری اطلاعات ۱۳۹۱/۱۰/۲۴.
- [9] A.M. Zare Bidoki, Designing and Implementing Crawl Systems and Ranking Persian Web Documents and Implementing a Persian Search Engine, 893222412, CyberSecurity Research Institute, IT Research Institute, 2013.
- [10] F. Dennis, N. Craswell and V. Vinay, "The impact of crawl policy on web search effectiveness." Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009.
- [11] "Web_crawler," [Online]. Available: www.en.wikipedia.org/wiki/Web_crawler.
- [12] "Apache Solr: List of security vulnerability-es," [Online]. Available: www.cvedetails.com/vulnerability-list/vendor_id-45/product_id-18263/Apache-Solr.html.
- [13] "Elasticsearch: Security vulnerabilities," [Online]. Available: www.cvedetails.com/vulnerabilitylist/vendor_id13554/Elasticsearch.html.
- [14] L. Jerri, Search Engine Optimization Bible, John Wiley & Sons,2009.
- [15] L. Theo, M. Brady and T. Masevic, "A Risk Assessment Method for Negative SEO Attacks



نسرین تاج مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرمافزار از دانشگاه آزاد اسلامی واحد تهران جنوب دریافت و همچنین کارشناسی ارشد خود را از دانشگاه علم و صنعت ایران در رشته

مهندسی فناوری ارتباطات و اطلاعات گرایش مخابرات امن دریافت کرد. ایشان هم‌اکنون دانشجوی دکترای DBA دانشگاه تهران است. زمینه‌های پژوهشی ایشان امنیت سامانه و اطلاعات، محروم‌گری هم‌چنین مدیریت استراتژیک و مدل‌های پیشرفته کسب و کار است. وی تاکنون مجری و ناظر چندین پروژه پژوهشی در پژوهشگاه فناوری ارتباطات و اطلاعات بوده است.



مهسا امیدوار سرکندي مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرمافزار در سال ۱۳۹۳

از دانشگاه آزاد اسلامی واحد تهران جنوب دریافت کرد. وی هم‌اکنون دانشجوی کارشناسی ارشد رشته نرمافزار دانشگاه صنعتی امیرکبیر است. زمینه‌های پژوهشی مورد علاقه ایشان امنیت نرم افزار، ارزیابی آسیب‌پذیری و تهدیدهای برنامه‌های کاربردی، تحلیل و طراحی و توسعه نرم افزارهای امن و پایگاههای داده‌های ابری است.

- [27] "Spamdexing" [Online]. Available: www.en.wikipedia.org/wiki/Spamdexing
- [28] "Enterprise Search" [Online]. Available: www.ideaeng.com/security-eprise-search-p1-0304
- [29] "Hadoop Framework" [Online]. Available: www.hadoop.apache.org.
- [30] "Heuristic Algorithm" [Online]. Available: [www.en.wikipedia.org/wiki/Heuristic_\(compute_r_science\)](http://www.en.wikipedia.org/wiki/Heuristic_(compute_r_science)).
- [31] "penguinoogle" [Online]. Available: www.mashable.com/2012/07/03/penguin-google-seo/#fGTDYX2hNaqX.
- [32] "How Search Works" [Online]. Available: www.google.com/insidesearch/howsearchworks/thetheory.
- [33] "How Search Works" [Online]. Available: www.support.google.com/webmasters/topic/6001971?hl=en&ref_topic=6001981.



حسن کوشکی مدرک کارشناسی خود را در رشته مهندسی فناوری اطلاعات در سال ۱۳۹۱ از دانشگاه پیام نور و مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات-امنیت اطلاعات در سال ۱۳۹۴ از دانشگاه تربیت مدرس دریافت کرد. زمینه پژوهشی مورد علاقه ایشان امنیت شبکه و اطلاعات، شبکه‌های کامپیوتری، امنیت مالتی مедیا، جریان‌سازی میدیا روی اینترنت، جریان‌سازی ویدیو نظیر به نظری، امنیت سامانه‌های مبتنی بر شهرت و طراحی امن معماری شبکه است.



شقایق نادری کارشناسی خود را در مهندسی کامپیوتر از دانشگاه خوارزمی در سال ۱۳۷۹ دریافت کرد. تحصیلات کارشناسی ارشد و دکترای خود را در رشته مهندسی کامپیوتر گرایش نرمافزار در دانشگاه تربیت مدرس به ترتیب در سال‌های ۱۳۸۱ و ۱۳۹۱ به پایان رسانده است. او در حال حاضر استادیار پژوهشگاه ارتباطات و فناوری اطلاعات ایران (مرکز تحقیقات مخابرات ایران) است. زمینه‌های پژوهشی مورد علاقه او درک تصویر، یادگیری ماشین، و امنیت وب است.