فصلنامه علمی - پژوهشی فرماندهی و کنترل، سال چهارم، شماره دو، تابستان ۱۳۹۹

فرماندهی و کنترل حساس به هزینه برای پایش محتوای وب با استفاده از عامل‌های خودمختار

سید مهدی نقیبی[۱*]، رضا انواری[۲]، علی فرقانی[۳]، بهروز مینایی[۴]

تاریخ دریافت: ۱۳۹۹/۰۲/۰۳ تاریخ پذیرش: ۱۳۹۹/۰۸/۱۳

چکیده

در این مقاله رویکردی چندعامله و حساس به هزینه برای فرماندهی و کنترل پایش محتوای وب انتخاب و کارکردهای کلیدی آن به‌طور کامل تشریح شده است؛ بر مبنای کارکردهای مذکور یک بیان رسمی برای فرماندهی و کنترل حساس به هزینه ارائه گردیده است. همچنین ساختار یک سامانه **C4ISR** مبتنی بر عامل‌های خودمختار پایشگر وب ارائه شده که در آن مؤلفه‌های هوشمند پردازش متن، پردازش تصویر، و خزش موضوعی به کار گرفته شده است. جهت مؤلفه پردازش تصویر، یک روش یادگیری حساس به هزینه برای شبکه‌های عصبی کانولوشن عمیق، و روشی جهت خزش موضوعی حساس به هزینه، پیشنهاد شده است. در روش پیشنهادی پردازش تصویر، زمانی که رده‌بندهای میانی که به ساختار **CNN** متعارف اضافه شده‌اند، به قطعیت لازم برای تعیین رده‌ی نمونه می‌رسند، فرآیند متوقف می‌شود، در غیر این صورت، رده‌بندی در لایه‌های بالاتر شبکه ادامه می‌یابد. به این ترتیب منابع پردازشی مدیریت و هزینه‌ها کاهش می‌یابد. در روش پیشنهادی خزش موضوعی به‌جای استفاده تنها از یک روش استخراج زمینه پیوند، از امتیازهای مجموعه‌ای از روش‌های استخراج زمینه پیوند برای افزایش کارایی استفاده شده است، که منجر به بهره‌برداری هدفمند از پهنای باند می‌گردد. نتایج آزمایش‌ها، نشان دهنده کارآمدی روش‌های پیشنهادی در مقایسه با بالاترین سطح نتایج موجود است. رویکرد حساس به هزینه ارائه شده در این مقاله، علاوه بر مسئله‌ی پایش محتوای وب، قابل به‌کارگیری در فرماندهی و کنترل سایر مسائل دنیای واقعی است.

واژگان کلیدی: پایش محتوای وب، رویکرد حساس به هزینه، رویکرد چند عامله، فرماندهی و کنترل.

[۱] دکتری مهندسی کامپیوتر- فرماندهی و کنترل، دانشگاه صنعتی مالک اشتر، مجتمع دانشگاهی برق و کامپیوتر (* مسئول مکاتبات) m.naghibi@chmail.ir

[۲] استاد مدعو، دانشگاه صنعتی مالک اشتر، مجتمع دانشگاهی برق و کامپیوتر rezaanvari@chmail.ir

[۳] استاد مدعو، دانشگاه صنعتی مالک اشتر، مجتمع دانشگاهی برق و کامپیوتر forghani@mut.ac.ir

[۴] دانشیار، دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر b_minaei@iust.ac.ir

۱. مقدمه

موضوع فضای سایبری و رهیافت‌های فرماندهی و کنترل برای پایش این فضا، در سال‌های اخیر همگام با توسعه‌ی بسیار سریع این حوزه، مورد توجه جدی محققان و سیاست‌گذاران قرار گرفته است [۱], [۲]. جوزف نای، نظریه‌پرداز برجسته، تقویت فرآیند پایش[1] در فضای سایبری را باعث افزایش قدرت بازدارندگی در این فضا می‌داند [۳]. فرماندهی و کنترل، و پایش فضای سایبری در سند راهبردی پدافند سایبری کشور [٤] بارها به‌عنوان موضوعی کلیدی مورد اشاره قرار گرفته است و پیاده سازی سامانه‌ای جامع که از قابلیت‌های اصلی آن پایش فضای سایبری است، به‌عنوان یکی از اهداف و راهبردها در سند مذکور معرفی شده است. از واژه **C4ISR**[2] برای ارجاع به فرماندهی و کنترل [٥] و همچنین سامانه‌هایی که فرماندهی و کنترل را پشتیبانی می‌کنند [٦]، استفاده می‌شود.

دیوید آلبرتز پژوهشگر برجسته فرماندهی و کنترل، داده‌های فراگیر[3] را یکی از چهار روند اصلی معرفی می‌کند که آینده **C4ISR** را تحت تأثیر قرار می‌دهد [۷]. منظور از داده‌های فراگیر، حجم بی‌سابقه از داده‌های خام و اطلاعات پردازش شده است که عامل‌های انسانی و سامانه‌های **C4ISR** با آن سر و کار دارند. از آنجا که در دنیای واقعی، منابع موجود برای پردازش داده‌ها محدود است، لازم است برای ایجاد چارچوب‌های فرماندهی و کنترل کارآمد، هزینه‌ی مربوط به فرآیندهای مدیریت و تحلیل اطلاعات در نظر گرفته شود. روش‌هایی که هزینه‌های مربوط به مراحل مختلف فرآیندها را در نظر می‌گیرند، روش‌های «حساس به هزینه[4]» نام دارند [۸]. گرچه روش‌ها و رویکردهای حساس به هزینه، موضوع پژوهش‌های مختلفی در حوزه یادگیری استقرایی را به خود اختصاص داده است اما در حوزه فرماندهی و کنترل تعداد کمی از پژوهش‌ها به این موضوع پرداخته‌اند [۹]–[۱۱] و به شکل مقتضی مورد بررسی قرار نگرفته است.

نظر به گستردگی بحث هزینه در فرماندهی و کنترل و اهمیت موضوع پایش فضای سایبری، جهت تعیین مرز مشخص برای پژوهش جاری و دستیابی به رهیافتی با قابلیت پیاده سازی عملی، در این مقاله یک رهیافت فرماندهی و کنترل حساس به هزینه برای پایش فضای وب ارائه شده است. طبق آمار ارائه شده در [۱۲] سرویس وب جزو پراستفاده‌ترین سرویس‌ها بر روی اینترنت از نظر حجم ترافیک اطلاعات محسوب می‌شود.

با توجه به خصوصیات فضای وب مانند توزیع‌شدگی و پویایی [۱]، استفاده از رویکردهای چندعامله[5] برای پایش وب یک بهره‌برداری مؤثر از این رویکرد محسوب می‌شود؛ به این ترتیب که مجموعه‌ای از عامل‌های هوشمند و خودمختار به‌صورت توزیع شده و در تعامل با یکدیگر به پایش محتوای وب می‌پردازند و محتویات مربوط به موضوعات هدف را شناسایی می‌کنند. رهیافت فرماندهی و کنترل متناظر با سامانه **C4ISR** پایش وب پیشنهادی این مقاله، در جایگاه سامان لبه‌ای از عامل‌های خودمختار، در فضای رهیافت‌های فرماندهی و کنترل قرار می‌گیرد. بر مبنای رهیافت فرماندهی و کنترل حساس به هزینه و چندعامله پیشنهادی، یک سامانه **C4ISR** پایش وب پیاده‌سازی و بر اساس معیارهای استاندارد، به شکل عملی مورد ارزیابی قرار گرفته است. روش‌های استفاده شده در پیاده‌سازی سامانه پایش محتوای وب عبارت‌اند از: پردازش متن، پردازش تصویر، و خزش موضوعی[6]. روش‌های مذکور هزینه‌های متفاوتی برای تحلیل محتویات وب در بر دارند؛ برخی پرهزینه‌تر و برخی کم‌هزینه‌تر هستند. این روش‌ها باعث هوشمندسازی سامانه پایش محتوا می‌شوند و از آن‌جا که مبتنی بر رویکردهای حساس به هزینه، طراحی و پیاده‌سازی شده‌اند، باعث افزایش سرعت محاسبات و ایجاد سامانه‌ای یکپارچه با قابلیت اقدام با کمترین تأخیر نسبت به عوامل محیطی می‌گردند.

در روش‌های پردازش متن، متون صفحات وب بررسی و صفحات حاوی محتویات هدف شناسایی می‌گردند [۱۳]. برای انجام این کار از انواع روش‌های رده‌بندی متن استفاده می‌شود. از جمله مهم‌ترین روش‌ها برای پایش وب و شناسایی محتویات

[1] Surveillance
[2] Command, Control, Communications, Intelligence, Surveillance, and Reconnaissance
[3] Ubiquitous data
[4] Cost-sensitive
[5] Multi-agent
[6] Topical crawling

هدف، تحلیل تصاویر موجود در صفحات وب است. برای انجام این کار می‌توان از انواع روش‌های یادگیری عمیق که در حال حاضر بهترین نتایج را در این حوزه به خود اختصاص داده‌اند استفاده نمود. هزینه‌ی لازم برای اجرای انواع روش‌های تحلیل تصاویر از متوسط تا بسیار زیاد، متغیر است. روش‌های یادگیری عمیق عموماً جزو روش‌های پرهزینه پردازش تصویر محسوب می‌شوند [۱٤]. در این مقاله یک روش یادگیری عمیق حساس به هزینه برای شبکه‌های عصبی کانولوشن[1] (**CNN**) ارائه شده است. در روش پیشنهادی تعدادی رده‌بند میانی و گره‌هایی با نام دروازه به ساختار **CNN** متعارف افزوده شده است. با دریافت هر نمونه‌ی ورودی، زمانی که رده‌بندهای میانی به اطمینان لازم برای تعیین رده‌ی نمونه می‌رسند، فرآیند متوقف می‌شود، در غیر این صورت، رده‌بندی در لایه‌های بالاتر شبکه ادامه می‌یابد. به این ترتیب با صرف هزینه کمتر برای نمونه‌های ساده و صرف هزینه بیشتر برای نمونه‌های پیچیده، منابع محاسباتی مدیریت می‌شود.

برای پایش فضای وب جهت شناسایی محتویات هدف به عامل‌هایی نرم‌افزاری نیاز است که صفحات را از وب دریافت و با استخراج پیوندهای موجود در این صفحات امکان کاوش سایر بخش‌های گراف وب را میسر سازند. این عامل‌های نرم‌افزاری خزشگر[2] نام دارند. طبق آمار [۱۵] در حال حاضر حدود ٦۰ میلیارد صفحه توسط موتورهای جستجو نمایه شده است؛ از این جهت برای پایش اطلاعات وب در زمان مطلوب راهی جز کاوش هدفمند این فضا و استفاده بهینه از پهنای‌باند وجود ندارد. یک راه حل برای مواجهه با این مشکل استفاده از روش‌های خزش موضوعی است. یک خزشگر موضوعی، خزشگری است که برای حرکت در زیرمجموعه‌ای از وب جهت شناسایی صفحات در رابطه با یک موضوع خاص می‌تواند پیوندهایی که منتهی به صفحات هدف می‌شوند را تعیین و از پیوندهایی که به صفحات نامربوط منتهی می‌شوند اجتناب نماید. در نتیجه با استفاده از حجم محدودی از منابع شبکه و پهنای باند، خزشگر موضوعی قادر خواهد بود صفحات با موضوع هدف را با پوششی قابل قبول از وب جمع‌آوری نماید [۱٦]. با به‌کارگیری روش‌های خزش موضوعی در یک سیستم پایش محتوای وب، صفحات حاوی محتویات هدف با سرعتی بیشتر و با مصرف حجم کمتری از پهنای باند مورد شناسایی قرار می‌گیرند. منظور از استخراج زمینه‌ی پیوند[3] در صفحه‌ی وب، شناسایی و استخراج واژه‌هایی از صفحه است که حوالی پیوند مورد بررسی واقع شده‌اند [۱۷]. زمینه‌ی پیوند شامل اشاره‌هایی به محتویات متناظر با آن است. با استخراج زمینه‌ی پیوندهای صفحات وب می‌توان موضوع محتویات متناظر با پیوندها را پیش‌بینی و موارد هدف را شناسایی نمود. در این مقاله یک روش پیشنهادی مبتنی بر استخراج زمینه پیوند برای هدایت بهتر خزشگرهای موضوعی ارائه شده است. ایده اصلی این روش بر مبنای ترکیب امتیاز کل صفحه و متن پیوند برای محاسبه امتیاز نهایی است [۱۸] با این تفاوت که به‌جای استفاده تنها از امتیاز متن پیوند، امتیاز مجموعه‌ای از روش‌های استخراج زمینه پیوند، شامل پنجره متن و بلوک برای افزایش کارایی خزشگرها استفاده شده است. به‌طور خلاصه سهم[4] علمی پژوهش جاری شامل چهار مورد زیر است:

- بررسی بحث هزینه در فرماندهی و کنترل، و به‌طور خاص **C4ISR** پایش وب، و ارائه یک بیان رسمی برای فرماندهی و کنترل حساس به هزینه، مبتنی بر کارکردهای کلیدی آن.
- ارائه ساختار یک سامانه **C4ISR** برای پایش وب برمبنای رویکرد حساس به هزینه و چندعامله با به‌کارگیری روش‌های پردازش متن و پردازش تصویر و خزش موضوعی.
- ارائه یک روش یادگیری حساس به هزینه برای **CNN**های عمیق بر مبنای رده‌بندها و دروازه‌های میانی در ساختار شبکه.
- معرفی یک روش پیشنهادی برای خزش موضوعی حساس به هزینه مبتنی بر ترکیب امتیاز روش‌های مختلف استخراج زمینه پیوند.

سامانه **C4ISR** پیشنهادی، شامل روش‌های پیشنهادی حساس به هزینه پردازش تصویر و خزش موضوعی، به شکل

[1] Convolutional Neural Network
[2] Crawler
[3] Link context extraction
[4] Contribution

عملی پیاده‌سازی و با استفاده از معیارهای شناخته شده، با سایر روش‌های موجود مقایسه شده است. روش پیشنهادی پردازش تصویر، قابل ترکیب با سایر روش‌های پیشرو در حوزه یادگیری عمیق است. همچنین روش پیشنهادی خزش موضوعی، در مقایسه با نتایج ارائه شده در [۱۸]، بالاترین سطح نتایج در این حوزه را به خود اختصاص داده است.

به‌طور مشخص، پژوهش حاضر تلاش دارد راه‌حلی برای در نظر گرفتن و مدیریت هزینه در فرماندهی و کنترل ارائه کند؛ مسئله‌ی پایش محتوای وب نیز به‌عنوان یک نمونه‌ی بااهمیت در حوزه فرماندهی و کنترل فضای سایبری، در قالب بستری برای پیاده‌سازی عملی بیان رسمی ارائه شده برای **C4ISR** حساس به هزینه، مورد بررسی قرار گرفته است. مقاله به این ترتیب ادامه می‌یابد: ابتدا پژوهش‌های مرتبط سه حوزه فرماندهی و کنترل سایبری، پردازش تصویر و خزش موضوعی، با تمرکز بر بحث هزینه در این پژوهش‌ها مورد بررسی و دسته‌بندی قرار گرفته است. سپس رهیافت مناسب فرماندهی و کنترل پایش وب، و کارکردهای کلیدی آن با رویکرد حساس به هزینه و چندعامله تبیین شده است. پس از آن سامانه **C4ISR** پایش وب معرفی و روش‌های پیشنهادی برای مؤلفه پردازش تصویر و خزش موضوعی حساس به هزینه تشریح شده است. نتایج مربوط به پیاده‌سازی همراه با نتیجه‌گیری در انتهای مقاله آمده است.

۲. پژوهش‌های مرتبط

۲-۱. هزینه در فرماندهی و کنترل سایبری

نویسندگان [۱۹] ضمن تشریح فرماندهی و کنترل گروه‌های حمله‌ی نیروی دریایی آینده، بحث هزینه در برخی اجزای سایبری سامانه مانند هزینه پهنای باند برای برقراری ارتباطات، هزینه فشرده‌سازی داده‌ها، و هزینه پردازش تصاویر جمع آورده شده از میدان نبرد را در سطح مفهوم مورد اشاره قرار داده‌اند ولی راهکار فنی برای آن ارائه ننموده‌اند. در [۲۰] راهکارهایی برای تقویت و یا جایگزینی یک سامانه موجود فرماندهی و کنترل نیروی دریایی آمریکا شامل نرم‌افزار و سخت‌افزار ارائه شده است. راهکارهای مذکور جهت دست‌یابی به سامانه‌ای مبتنی بر نرم‌افزار که از قابلیت ایجاد آگاهی موقعیتی و تصویر عملیاتی مشترک کارآمدتر برخوردار است، ارائه شده است. مقایسه هزینه‌ی روش‌های نرم‌افزاری مختلف، بر اساس اندازه نرم‌افزار و میزان امکان استفاده مجدد از کدها، با تخمینی از تعداد خطوط کد، انجام شده است. تحلیل کلان داده‌ها[1] در سامانه‌های فرماندهی و کنترل توسط [۱۱] مورد بررسی قرار است و استفاده از روش‌های پردازش توزیع شده برای حل چالش‌های آن پیشنهاد شده است. نویسندگان [۱۰] بحث هزینه را در سامانه‌های پایش توده[2] در مقیاس بالا مورد بررسی قرار داده‌اند و با ارائه یک تحلیل ریاضی بر اساس تناقض مثبت نادرست[3]، با استفاده از آزمون فرضیه‌های دودویی، تروریست یا بی‌گناه بودن افراد را تعیین نموده‌اند. هدف نهایی این سامانه، کاهش هزینه کلی فرآیند پایش توده است.

در [۲۱]، از یک دیدگاه متفاوت به هزینه‌هایی که طرح‌های پایش اطلاعات آژانس امنیت ملی آمریکا[4] (**NSA**) به اقتصاد شرکت‌ها، سیاست خارجی، و فناوری‌های مورد استفاده در فضای اینترنت، تحمیل می‌کند پرداخته شده است. هزینه‌ی محلی‌سازی داده‌ها[5] و مرزبندی اینترنت، از جمله هزینه‌های مربوط به فضای سایبری است که در [۲۱] به ارتباط افزایش این هزینه‌ها با پایش فضای سایبری پرداخته شده است.

با بررسی پژوهش‌های پیشین درمی‌یابیم که بحث هزینه در حوزه فرماندهی و کنترل و به‌طور خاص **C4ISR** سایبری به شکل مقتضی مورد مطالعه قرار نگرفته است؛ اما این موضوع در حوزه یادگیری استقرایی، موضوع پژوهش‌های فراوانی بوده است. در [۸] یک طبقه‌بندی نسبتاً جامع از هزینه‌های مختلفی که ممکن است در یادگیری استقرایی مفاهیم واقع شوند، ارائه شده است. مفهوم «هزینه» در این مقاله، به انتزاعی‌ترین شکل ممکن در نظر گرفته شده است؛ مثلاً در پردازش تصویر، هزینه می‌تواند در قالب زمان **CPU** مورد نیاز برای انجام محاسبات لازم، اندازه‌گیری شود. نویسنده [۸] هزینه‌های ممکن در یادگیری استقرایی مفاهیم را به ۹ دسته تقسیم می‌کند که

[1] Big data
[2] Mass surveillance
[3] False-positive paradox
[4] National Security Agency
[5] Data localization

عبارت‌اند از: هزینه‌ی خطاهای رده‌بندی نادرست، هزینه آزمون‌ها، هزینه محاسبات، هزینه مربی[1]، هزینه نمونه‌ها[2]، هزینه تعامل انسان و کامپیوتر[3] (**HCI**)، هزینه مداخله[4]، هزینه دست‌آوردهای ناخواسته[5]، و هزینه ناپایداری[6]. با توجه به ماهیت سایبری سامانه **C4ISR** پایش وب، با بهره‌گیری از انواع هزینه‌های احصا شده برای فرآیند استقرای مفاهیم، در این مقاله یک بیان رسمی بر مبنای کارکردهای کلیدی فرماندهی و کنترل، و به‌طور خاص **C4ISR** پایش وب ارائه خواهیم نمود.

۲-۲. پردازش تصویر حساس به هزینه

هزینه محاسباتی یک چالش واقعی برای **CNN**های عمیق است. روش‌های موجود برای کاهش هزینه **CNN**ها به سه دسته تقسیم می‌شوند: اصلاح مدل موجود، استفاده از محاسبات پیشرفته و سطح پایین، و روش‌های وفقی[7].

۲-۲-۱. اصلاح مدل موجود

دسته اول شامل روش‌هایی برای ایجاد مدل‌های تخمینی است که با یادگیری مدل‌های جدید یا اصلاح[8] مدل‌های موجود هزینه را کاهش می‌دهند [۲۲]. روش تقلید[9] شبکه، با آموزش یک «شبکه کم عمق» [۲۳] یا یک «شبکه مناسب[10]» [۲۴]، یک مدل جدید آموزش می‌دهد که مدل شاگرد[11] نام دارد. این مدل جدید از بنیان ایجاد می‌شود تا رفتار مدل اصلی که مدل معلم نامیده می‌شود را تقلید کند. تجزیه‌ی شبکه [۲۵] نوع دیگری از تخمین است.

۲-۲-۲. استفاده از محاسبات پیشرفته و سطح پایین

این روش‌ها بدون این که ساختار شبکه را اصلاح کنند، سرعت پردازش شبکه را افزایش می‌دهند. یک خانواده از این روش‌ها، نحوه انجام محاسبات لایه‌ها را هدف قرار می‌دهد و برای انجام آن از کانولوشن‌های مبتنی بر تبدیل سریع فوریه استفاده می‌کنند [۲۶]. خانواده دیگری از این گروه، میزان بهره‌گیری از سخت‌افزار را افزایش می‌دهند [۲۷]، [۲۸]؛ مانند موازی‌سازی سطح پایین، استفاده مؤثر از حافظه، و انجام محاسبات ریاضی با دقت پایین.

۲-۲-۳. روش‌های وفقی

بر خلاف هر دو روش قبلی که دارای رفتار ایستا با تمام نمونه‌های ورودی هستند، روش‌های تطبیقی، منابع محاسباتی را با یک سیاست وابسته به ورودی، اختصاص می‌دهند. روش‌های تطبیقی می‌توانند با دو دسته روش قبلی ترکیب شوند و از مزایای هر دو استفاده کنند.

«شبکه تصمیم گیری عمیق» [۲۹] میزان دشواری نمونه‌ها را تشخیص می‌دهد و تصاویر مشکل‌تر را به مدل‌های بعدی در ساختار آبشار[12] منتقل می‌کند. روش «آبشار شبکه‌های عصبی کانولوشن» [۳۰] بر روی چند تفکیک‌پذیری[13] متفاوت تصویر عمل می‌کند، به‌سرعت نواحی پس‌زمینه را در مراحل با تفکیک‌پذیری پایین کنار می‌گذارد، و تعداد محدودی کاندید چالش‌برانگیز را در مراحل آخر با تفکیک‌پذیری بالا ارزیابی می‌کند. روش «حالت عمیق[14]» [۳۱] با استفاده از استراتژی تقسیم و غلبه، یک چارچوب رگرسیون آبشاری عمیق برای تخمین حالت انسان ارائه می‌کند.

لی و همکاران [۱۴] با یک روش آبشار متفاوت از پژوهش‌های پیشین، روشی به نام «آبشار لایه‌ای عمیق[15]» برای مسئله تقسیم‌بندی معنایی تصویر ارائه نموده‌اند. آبشار لایه‌ای، برخلاف آبشارهای مدل که از مجموعه‌ای از مدل‌ها استفاده می‌کنند، یک شبکه واحد با چند شاخه‌ی میانی آموزش می‌دهد. فرآیند تقسیم‌بندی معنایی از شاخه‌های پایین‌تر آغاز می‌شود و این شاخه‌ها درجه اطمینان‌هایی برای هر تصویردانه[16] محاسبه می‌کنند. این روند برای تصویردانه‌های آسان‌تری که در لایه‌های پایین شبکه شناسایی شده‌اند، متوقف می‌شود و نواحی دشوارتر به سطوح بالاتر شبکه عمیق ارجاع می‌گردند. روش ارائه شده در مقاله‌ی حاضر مشابه روش آبشار لایه‌ای عمیق است، اما به‌جای استفاده از این روش برای تعیین میزان دشواری نواحی

[1] Teacher
[2] Cost of cases
[3] Human Computer Interaction
[4] Intervention
[5] Unwanted achivements
[6] Instability
[7] Adaptive
[8] Modify
[9] Mimicking
[10] Fitnet
[11] Student
[12] Cascade
[13] Resolution
[14] DeepPose
[15] Deep layer cascade
[16] Pixel

مختلف تصویر، از آن برای تعیین دشواری کل تصویر جهت فرآیند رده‌بندی و شناسایی موضوع تصویر استفاده شده است.

۲-۳. خزش وب حساس به هزینه

در حوزه‌ی خزشگرهای موضوعی پژوهش‌های زیادی انجام شده است. به‌عنوان نمونه در [۳۲] و [۳۳] امکان استفاده از خزشگرهای موضوعی بر روی آرشیوهای موجود وب و نمایه‌ی موتور جستجو، به ترتیب برای ایجاد مجموعه‌هایی از رویدادها و جمع‌آوری محتویات علمی، باهدف کاهش زمان خزش بررسی شده است. نویسندگان [۳٤] و [۳۵] از روش‌های یادگیری برای ایجاد بهبود در سیاست انتخاب پیوندها جهت خزش استفاده نموده‌اند. همچنین در [۳٦] از خزش موضوعی برای ردگیری[1] خودکار رویدادها و آرشیو نمودن آن‌ها استفاده شده است. یک گروه اصلی از روش‌های خزش موضوعی از روش‌های استخراج زمینه پیوند برای بهینه‌سازی استفاده از پهنای باند استفاده می‌کنند. طبق نتایج [۱۸] این دسته از روش‌ها بالاترین میزان کارایی را به خود اختصاص داده‌اند؛ از این جهت در ادامه به دسته‌بندی روش‌های خزش موضوعی بر مبنای روش شناسایی زمینه پیوند می‌پردازیم.

۲-۳-۱. متن صفحه و پیوند

آسان‌ترین راه برای استخراج زمینه پیوند، در نظر گرفتن کل متن یک صفحه وب به‌عنوان زمینه پیوند همه پیوندهای صفحه است. خزشگر ارائه شده در [۳۷] که «جستجوی ماهی» نامیده شده است، از این روش برای ارزیابی پیوندهای یک صفحه وب استفاده نموده است. روش اول بهترین[2] که یک نسخه اصلاح شده از این الگوریتم است، امتیاز متن صفحه را با امتیاز متن پیوند ترکیب می‌کند. این ترکیب بر اساس فرمول زیر صورت می‌گیرد:

$$\begin{aligned} &\boldsymbol{link\_score} \\ &= \boldsymbol{\beta} \times \boldsymbol{Relevancy}(\boldsymbol{page\_text}) + (\mathbf{1}-\boldsymbol{\beta}) \\ &\times \boldsymbol{Relevancy}(\boldsymbol{link\_context}) \end{aligned} \qquad (۱)$$

که در آن $\boldsymbol{link\_score}$ امتیاز پیوند درون صفحه، $\boldsymbol{page\_text}$ متن کل صفحه، و $\boldsymbol{link\_context}$ زمینه پیوند استخراج شده از پیوند است که در یک روش اول بهترین معمولی، برابر با متن پیوند در نظر گرفته می‌شود. تابع $\boldsymbol{Relevancy}$ وظیفه محاسبه امتیاز یک متن معین را بر اساس موضوع هدف بر عهده دارد.

۲-۳-۲. روش پنجره متن

این روش مؤثر، تعداد $\boldsymbol{W}$ کلمه مجاور یک پیوند را به‌عنوان زمینه پیوند در نظر می‌گیرد [۱۷]. برای این کار از یک پنجره متقارن در اطراف پیوند استفاده می‌شود و تعداد $\boldsymbol{W/2}$ کلمه قبل و تعداد $\boldsymbol{W/2}$ کلمه پس از متن پیوند استخراج می‌گردد. همچنین متن پیوند داخل پنجره قرار داده می‌شود. چالش حل نشده این روش، نامعلوم بودن تعداد مطلوب کلمات برای استخراج زمینه پیوند است.

۲-۳-۳. روش بلوک

این روش‌ها برای یافتن زمینه‌ی یک پیوند از اطلاعات بلوکی که پیوند در آن واقع شده است، استفاده می‌نماید. بلوک‌ها نواحی تقریباً مستطیل شکلی از صفحه هستند که دارای محتوای نزدیک به هم هستند. الگوریتم قطعه‌بندی مبتنی بر منظره‌ی صفحه[3] (**VIPS**) یکی از بهترین روش‌ها برای استخراج زمینه‌ی پیوند بر اساس بلوک‌بندی صفحات است [۳۸]. این الگوریتم سه مرحله دارد: استخراج بلوک‌ها، شناسایی تفکیک کننده‌ها و ساختن ساختار محتوا. این سه مرحله بر روی هم به‌عنوان یک دور[4] از اجرای رویه محسوب می‌شوند.

۳. تبیین فرماندهی و کنترل پایش وب

دیوید آلبرتز در [۵] شش کارکرد کلیدی[5] برای فرماندهی و کنترل برمی‌شمرد. این کارکردها عبارت‌اند از: دیدبانی فضای نبرد[6]، آگاهی[7]، درک موقعیت نظامی[8]، فرآیند حس‌سازی[9]، توسعه‌ی قصد فرمان[10]، و مدیریت فضای نبرد[11]. تبیین این کارکردها برای رویکرد مورد استفاده، ویژگی‌های **C4ISR** مورد نظر برای پایش وب را تعیین می‌کند. در این مقاله، رویکردی حساس به هزینه و چند عامله برای فرماندهی و کنترل پایش

[1] Tracking
[2] Best-first
[3] Vision Based Page Segmentation Algorithm
[4] Round
[5] Key function
[6] Battlespace monitoring
[7] Awareness
[8] Understanding the military situation
[9] The sensemaking process
[10] Developing command intent
[11] Battlespace management

وب پیشنهاد شده است. در ادامه، به دلایل انتخاب رویکرد چند عامله برای این مسئله‌ی می‌پردازیم، سپس با تشریح انواع هزینه‌ها در فرآیند فرماندهی و کنترل سایبری، شش کارکرد کلیدی را برای رویکرد حساس به هزینه و چند عامله پیشنهادی تبیین می‌کنیم.

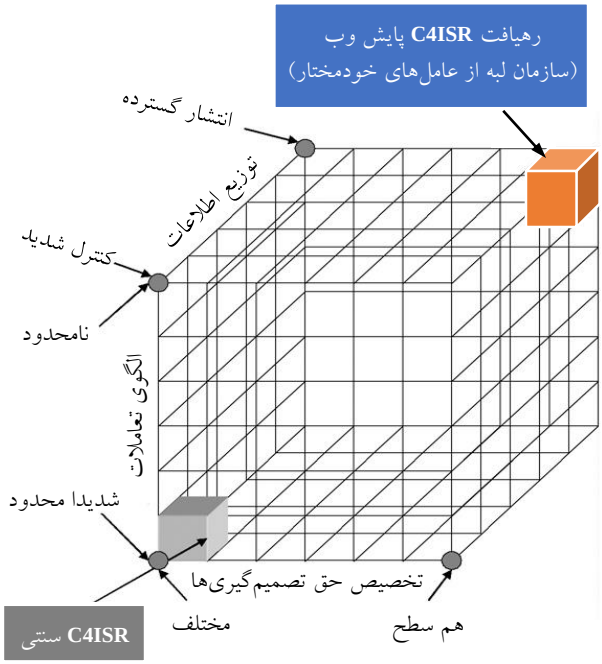
۳-۱. استفاده از رویکرد چند عامله

موجودیت‌های سایبری که منبع تولید قدرت سایبری جهت پایش وب محسوب می‌شوند، در واقع عامل‌هایی خودمختار هستند که با بر عهده گرفتن وظیفه پایش، اقدام به گشت و گزار در فضای وب می‌کنند و منابع هدف را در آن شناسایی می‌نمایند. ما از این موجودیت‌های نرم‌افزاری با عنوان «عامل‌های پایشگر» یاد می‌کنیم.

شرایط حاکم بر محیط عملکرد عامل‌های پایشگر یعنی وب و همچنین ماهیت این موجودیت‌های خودمختار، استفاده از رهیافت‌های چند عامله را برای طراحی و پیاده‌سازی آن‌ها کاملاً توجیه می‌نماید. از آنجا که بخشی از کارکردهای فرماندهی و کنترل در دامنه‌ی شناخت انجام می‌پذیرد [٦]، لازم است عامل‌های پایشگر از توانایی نگاشت عناصر موجود در دامنه‌ی اطلاعات به دامنه‌ی شناخت برخوردار باشند. در بخش ارائه ساختار فرماندهی و کنترل پایش وب، راه‌حل‌هایی برای نحوه‌ی ایجاد هماهنگی و تقسیم وظایف بین عامل‌ها ارائه خواهد شد. با توجه به ماهیت مستقل عامل‌ها، ساختار پیشنهادی کاملاً مقیاس‌پذیری[1] است.

۳-۲. رهیافت فرماندهی و کنترل مناسب پایش وب

آلبرتز در [٦]، برای فضای رهیافت‌های فرماندهی و کنترل، سه بعد معرفی می‌کند. در واقع این ابعاد، سه ویژگی اصلی هستند که جوهره فرماندهی و کنترل را تعیین می‌نمایند. این سه عبارت‌اند از: تخصیص حق تصمیم‌گیری‌ها، الگوهای تعاملات بین کنشگرها[2]، و توزیع اطلاعات. شکل ۱ جایگاه رهیافت **C4ISR** پایش وب را در فضای رهیافت‌های فرماندهی و کنترل نمایش می‌دهد. همان‌طور که در شکل مشاهده می‌شود، هر اندازه که حق تصمیم‌گیری‌ها هم‌سطح‌تر، تعاملات بین کنشگرها نامحدودتر، و توزیع اطلاعات به شکلی گسترده‌تر انجام شود، رویکرد فرماندهی و کنترل ایجاد شده، سازمان عمل‌کننده را به یک سازمان لبه، نزدیک‌تر می‌کند.



شکل ۱. فضای رهیافت‌های فرماندهی و کنترل [٦] و نمایش جایگاه رهیافت C4ISR پایش وب در آن.

فضای وب، ویژگی‌هایی دارد که استفاده از رهیافت‌های فرماندهی و کنترل سنتی برای پایش آن را ناکارآمد می‌کند. ویژگی‌هایی مانند گستردگی، ناهمگنی، توزیع شدگی و پویایی فضای وب، به‌روشنی بیان‌کننده‌ی این مسئله است که برای پایش محتویات وب در این فضا به رهیافت فرماندهی و کنترلی نیاز است که عامل‌ها را هر چه بیشتر به یک سازمان لبه نزدیک کند. به دلیل ماهیت سایبری عامل‌های پایشگر، امکان برقراری ارتباط و توزیع اطلاعات بین آن‌ها با استفاده از یک بستر ارتباط محلی با هزینه بسیار کم فراهم است. از این رو می‌توان با بهره گیری از رهیافت‌های فرماندهی و کنترل نزدیک به سازمان لبه، رویکردی حساس به هزینه و چند عامله‌ای ایجاد کرد که در آن همه‌ی عامل‌های پایشگر از حق تصمیم‌گیری یکسانی برخوردار باشند، امکان تعامل بین عامل‌ها به‌طور کامل فراهم باشد و همه‌ی آن‌ها به اطلاعات به دست آمده توسط سایرین دسترسی داشته باشند. به این ترتیب همه‌ی عامل‌ها قادر خواهند بود با بهره‌گیری

[1] Scalability

[2] Actors

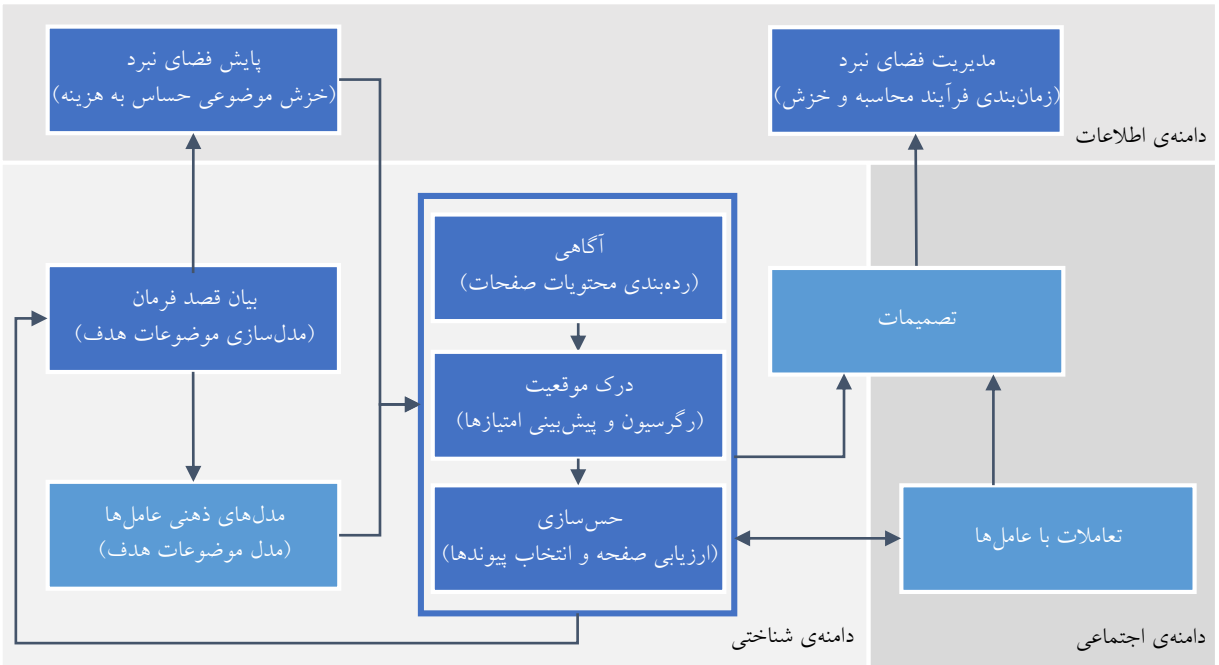
از اطلاعات به دست آمده توسط سایرین، بهترین تصمیم را اتخاذ و بالاترین میزان کارایی را از خود بروز دهند.

۳-۳. کارکردهای فرماندهی و کنترل برای رویکرد چند عامله پایش وب

در این بخش، شش کارکرد کلیدی فرماندهی و کنترل معرفی شده توسط آلبرتز در [۵] را با رویکردی چند عامله برای پایش وب تشریح می‌نماییم. رویه اتخاذ شده برای تبیین کارکردها، مشابه رویه استفاده شده در [۳۹] جهت تشریح کارکردهای فرماندهی و کنترل ارتشی از اشیاء هوشمند است. شکل ۲ ارتباط بین این شش کارکرد را نشان می‌دهد. با توجه به ماهیت سایبری و چندعامله مسئله پایش محتوای وب، فرآیند مذکور در دامنه‌های اطلاعات، شناختی، و اجتماعی جریان دارد. همان‌طور که مشاهده می‌شود با بیان قصد فرمان در قالب مدل‌های استقرا شده از موضوع و ارائه آن به مدل‌های ذهنی عامل‌های پایشگر، فرآیند خزش و دیدبانی فضای نبرد آغاز و با تحلیل محتوای هر صفحه دریافت شده، و رده‌بندی و تعیین امتیاز پیوندهای داخل آن و ارزیابی پیش‌بینی انجام شده در مورد موضوع آن سه کارکرد آگاهی، درک موقعیت و حس‌سازی انجام می‌شود. طی تعامل عامل‌ها با یکدیگر و تبادل اطلاعات مربوط به صفحات و پیوندهای پایش شده، تصمیمات نهایی برای پایش صفحات جدید بین عامل‌ها به اشتراک گذاشته می‌شود و کارکرد مدیریت فضای نبرد، زمان‌بندی و برنامه ریزی اجرای تصمیمات را بر عهده می‌گیرد. همچنین این امکان وجود دارد که با انتقال بازخورد از فرآیند طی شده به کارکرد قصد فرمان، بیان صحیح‌تری از مدل‌های بازنمایی کننده‌ی قصد فرمان ایجاد شود و در اختیار عامل‌های پایشگر قرار گیرد. در ادامه به توضیح بیشتر این کارکردها برای پایش وب می‌پردازیم.

۳-۳-۱. دیدبانی فضای نبرد

دیدبانی فضای نبرد شامل جمع‌آوری اطلاعات از همه‌ی جنبه‌های فضای نبرد و پردازش اولیه‌ی این اطلاعات است [۵]. فضای نبرد در مسئله پایش وب، خدمات وب اینترنت است. از این جهت، فرآیند پایش وب در دامنه‌ی اطلاعات انجام می‌شود. عامل‌های پایشگر، بر مبنای این کارکرد، وظیفه دارند با کاوش در فضای سایبری، اطلاعاتی مانند ساختار پیوند بین صفحات وب و محتوای صفحات را از فضای اینترنت جمع‌آوری کنند. همچنین بر مبنای کارکرد دیدبانی فضای نبرد، عامل‌های پایشگر وظیفه دارند پردازش‌های اولیه را برای آماده سازی این اطلاعات جهت استفاده در فرآیند دست‌یابی به آگاهی، انجام دهند. اطلاعات به دست آمده توسط هر یک عامل‌های پایشگر در اختیار سایر عامل‌ها نیز قرار می‌گیرد.



شکل ۲. ارتباط بین شش کارکرد کلیدی فرماندهی و کنترل پایش محتوای وب.

۳-۳-۲. آگاهی

«آگاهی» یا «آگاهی موقعیتی» عبارت است از دریافت یک عامل از اطلاعات درباره‌ی موقعیت. آگاهی شامل ترکیبی از دانش پیشین و اطلاعات جاری است. در واقع آگاهی، همبسته‌ی شناختی اطلاعات است [۵]. در مسئله پایش وب، عامل‌های پایشگر حین مواجهه با یک صفحه‌ی وب با ترکیب دانش قبلی خود در مورد محتویات هدف، و اطلاعات استخراج شده از نمونه‌ی جاری و سایر بخش‌های فضای وب، در رابطه با نمونه‌ی جاری به آگاهی دست پیدا می‌کند. این فرآیند با بهره‌گیری از قابلیت نگاشت اطلاعات به آگاهی توسط عامل پایشگر انجام می‌گیرد. دانش قبلی عامل‌ها در قالب مدل‌هایی برای شناسایی محتویات هدف نگه‌داری می‌شود. این مدل‌ها، رده‌بندهایی هستند که توسط نمونه‌های هدف موجود آموزش دیده‌اند و می‌توانند با دریافت نمونه‌های جدید بروز رسانی شوند. آگاهی به‌دست آمده می‌تواند بین همه‌ی عامل‌ها به اشتراک گذاشته شود.

۳-۳-۳. درک موقعیت

آگاهی و دانش، ورودی‌های فرآیند ادراک هستند. این کارکرد، پیش‌بینی رویدادها و شناسایی عدم قطعیت‌ها را شامل می‌شود [۵]. در مسئله پایش وب، بر مبنای این کارکرد، عامل‌های پایشگر پس از کسب آگاهی در مورد صفحات وب پایش شده موجود در فضای اینترنت، با یک دید عمیق‌تر موارد غیرقطعی مرتبط با نمونه‌های جاری و سایر بخش‌های فضای وب را تعیین می‌کنند. عامل‌های پایشگر باید پیش‌بینی کنند که اولاً محتویات صفحات جاری با چه احتمالی به موضوعات هدف مرتبط هستند، ثانیاً دنبال نمودن پیوندهای موجود در صفحات جاری با چه قطعیتی به صفحات هدف منتهی می‌شوند.

۳-۳-٤. حس‌سازی

عامل‌ها در فرآیند حس‌سازی به دنبال انجام سه فعالیت به هم مرتبط هستند: تولید مجموعه‌ای از عمل‌های جایگزین به‌منظور کنترل موقعیت، شناسایی معیارهایی که بتوان با استفاده از آن‌ها عمل‌های جایگزین را باهم مقایسه نمود، و انجام ارزیابی این عمل‌ها. این فرآیند در دامنه‌های شناختی و اجتماعی اتفاق می‌افتد. تصمیمات، محصول فرآیند حس‌سازی هستند [۵].

عامل‌های پایشگری که جهت شناسایی محتویات هدف در فضای وب به کار گرفته می‌شوند، همواره در حال تولید مجموعه‌ای از عمل‌های جایگزین برای انجام، و ارزیابی این عمل‌ها بر اساس معیارهای مختلف هستند. عامل‌ها با دریافت هر صفحه جدید و استخراج پیوندهای آن، با فراهم نمودن امکان دنبال کردن این پیوندها، مجموعه‌ای از عمل‌های جایگزین در اختیار مجموعه‌ی عامل‌ها قرار می‌دهند و با محاسبه‌ی ارزش دنبال نمودن هر یک از این پیوندها توسط روش‌های خزش موضوعی، تصمیم می‌گیرند که کدام پیوند انتخاب شود. پس از دریافت هر صفحه وب، با استفاده از رده‌بندهای ارزیاب یا عامل انسانی، عمل انجام شده، ارزیابی می‌شود. بازخوردهای به دست آمده، می‌تواند برای تصحیح مدل‌های آموخته شده استفاده شود.

۳-۳-۵. توسعه قصد فرمان

«قصد»، بیانی از هدف است. قصد فرمان، محصولی است که توسط فرآیند فرماندهی و کنترل تولید می‌شود [٦]. در مسئله‌ی پایش محتوای وب، قصد فرمان، شناسایی محتویات هدف با هر شکل و قالب در فضای وب است. بحث اصلی، نحوه‌ی بازنمایی و بیان این قصد است. در واقع باید از قالب‌ها و الگوهای گوناگون محتویات هدف بازنمایی‌هایی ایجاد شود که عامل‌های پایشگر را قادر به درک قصد فرمان، یعنی شناسایی و پایش محتویات هدف نماید. از جمله روش‌هایی که می‌تواند برای مدل‌سازی و بازنمایی محتویات هدف به کار گرفته شود، روش‌های رده‌بندی[1] است. این روش‌ها با دریافت مجموعه‌ای از نمونه‌های متعلق به رده‌های[2] مختلف، می‌توانند مدلی کنند که قادر است رده‌ی مربوط به نمونه‌های جدید را شناسایی کند. روش‌های رده‌بندی امکان شناسایی محتویات هدف با قالب‌های متفاوتی مثل متن و تصویر را فراهم می‌کند.

۳-۳-٦. مدیریت فضای نبرد

هدف این فرآیند انعکاس قصد فرمان در قالب یک طرح[3]، انتشار سریع طرح، نظارت بر اجرا و تشخیص به‌موقع نیاز به

[1] Classification
[2] Class
[3] Plan

تغییرات در آن است. در طرح، این موارد مشخص می‌شود: کاری که باید انجام شود، منابع مورد استفاده، زمان‌بندی، محل انجام، و شرایط تغییر چهار مورد پیشین [۵]. عامل‌های پایشگر در تعامل با یکدیگر اطلاعات صفحات پایش شده شامل امتیاز محاسبه شده برای پیوندها را تبادل می‌کنند و تصمیمات مربوط به بهترین پیوندهای انتخاب شده را به اشتراک می‌گذارند. طی کارکرد مدیریت فضای نبرد در مسئله پایش محتوای وب، اختصاص منابع محاسباتی و پهنای باند لازم برای دریافت و تحلیل پیوندهای هدف انجام می‌شود و زمان‌بندی اجرای فرآیند به‌صورت موازی برای مجموعه‌ای از عامل‌های پایشگر صورت می‌پذیرد.

۳-٤. فرماندهی و کنترل حساس به هزینه

در این مقاله از مدلی مشابه با آنچه در [٤٠] بر مبنای سه مرحله اصلی فرآیند رده‌بندی ارائه شده است، برای مدل سازی هزینه فرآیند فرماندهی و کنترل استفاده می‌شود. در ادامه تعریفی رسمی برای مقایسه سودمندی دو فرآیند فرماندهی و کنترل، ارائه می‌دهیم.

تعریف: فرآیند فرماندهی و کنترل $\boldsymbol{A}$ از فرآیند فرماندهی و کنترل $\boldsymbol{B}$ سودمندتر است، اگر و فقط اگر:

$$\boldsymbol{Cost_{total}(A) < Cost_{total}(B)} \qquad (۲)$$

که در آن $\boldsymbol{Cost_{total}}$، هزینه کل و برابر با جمع همه‌ی هزینه‌های انجام شده در مراحل مختلف فرآیند فرماندهی و کنترل است. هزینه کل می‌تواند با استفاده از رابطه‌ی زیر بر مبنای شش کارکرد کلیدی فرماندهی و کنترل محاسبه شود:

$$\begin{aligned}\boldsymbol{Cost_{total}(P)} = &\boldsymbol{Cost_{bs\_monitoring}(P)} \\ &+ \boldsymbol{Cost_{awareness}(P)} \\ &+ \boldsymbol{Cost_{understanding\_ms}(P)} \\ &+ \boldsymbol{Cost_{sensmaking}(P)} \\ &+ \boldsymbol{Cost_{command\_intent}(P)} \\ &+ \boldsymbol{Cost_{bs\_management}(P)}\end{aligned} \qquad (۳)$$

که در آن $\boldsymbol{Cost_{bs\_monitoring}}$ هزینه دیدبانی فضای نبرد، $\boldsymbol{Cost_{awareness}}$ هزینه آگاهی، $\boldsymbol{Cost_{understanding\_ms}}$ هزینه درک موقعیت، $\boldsymbol{Cost_{sensmaking}}$ هزینه حس‌سازی، $\boldsymbol{Cost_{command\_intent}}$ هزینه توسعه قصد فرمان، و $\boldsymbol{Cost_{bs\_management}}$ هزینه مدیریت فضای نبرد است.

در واقع مسئله‌ی که باید حل شود کمینه‌سازی هزینه فرآیند فرماندهی و کنترل بر اساس رابطه ۳ است، و در صورت انجام این کار، فرماندهی و کنترل با سودمندی/هزینه بهینه حاصل می‌شود.

۳-۵. رویکرد حساس به هزینه برای کارکردهای فرماندهی و کنترل پایش وب

در این بخش با یک رویکرد حساس به هزینه، شش کارکرد کلیدی فرماندهی و کنترل را برای پایش وب بررسی و هزینه‌های مربوط به هر کارکرد را تعیین می‌کنیم. لازم به ذکر است هزینه‌هایی مانند محاسبات در همه‌ی این کارکردها وجود دارد؛ اما در برخی از آن‌ها در مقایسه با سایر هزینه‌ها بسیار ناچیز است و در برخی دیگر بخش اصلی هزینه‌ها را به خود اختصاص می‌دهد. از این جهت برای شفاف‌تر شدن بحث، تنها به هزینه‌های اصلی هر کارکرد پرداخته شده است. جدول ۱ هزینه مربوط به کارکردهای فرماندهی و کنترل را برای پایش وب نشان می‌دهد. در ادامه به تشریح هر یک از این هزینه‌ها پرداخته می‌شود.

جدول ۱. هزینه‌های اصلی کارکردهای فرماندهی و کنترل پایش وب

کارکرد **C4ISR**	هزینه‌های اصلی
دیدبانی فضای نبرد	پهنای باند
آگاهی	خطای رده‌بندی نادرست، آزمون‌های زمان آزمایش، محاسبات زمان آزمایش
درک موقعیت نظامی	خطا در فرآیند رگرسیون
فرآیند حس‌سازی	**HCI** ارزیابی پیش‌بینی‌ها، محاسبات ارزیابی پیش‌بینی‌ها
و توسعه‌ی قصد فرمان	نمونه‌ها، مربی، آزمون‌های زمان آموزش، محاسبات زمان آموزش
مدیریت فضای نبرد	محاسبات برنامه‌ریزی پایشگرها

۳-۵-۱. هزینه دیدبانی فضای نبرد

فضای نبرد پایش وب، در دامنه‌ی اطلاعات قرار دارد و عامل‌ها برای کاوش این فضا و دسترسی به محتوای آن بر روی

یک خط ارتباط با اینترنت عمل می‌کنند. بنا بر این هزینه اصلی در این کارکرد، پهنای‌باند مورد استفاده توسط عامل‌های پایشگر است. در این مقاله از روش‌های خزش موضوعی برای دیدبانی فضای وب استفاده شده است. به بیان دیگر برای فرآیند فرماندهی و کنترل پایش وب $\boldsymbol{W}$ خواهیم داشت:

$$\boldsymbol{Cost_{bs\_monitoring}(W) = Cost_{bandwidth\_usage}(W)} \quad (٤)$$

که در آن $\boldsymbol{Cost_{bandwidth\_usage}}$ هزینه استفاده از پهنای باند است. همچنین در این کارکرد هزینه‌هایی صرف پردازش اولیه اطلاعات صفحات وب می‌شود که در قیاس با سایر هزینه‌ها قابل چشم پوشی است.

۳-۵-۲. هزینه آگاهی

در این کارکرد مدل‌های از پیش آموزش دیده که همان رده‌بندها هستند توسط عامل‌های پایشگر برای رسیدن به آگاهی مورد استفاده قرار می‌گیرند و هزینه‌های اصلی در این کارکرد نیز مربوط به استفاده از این مدل‌ها است. هزینه محاسباتی زمان آزمایش، هزینه آزمون‌ها، و هزینه رده‌بندی نادرست مواردی هستند که در کارکرد آگاهی برای پایش وب باید در نظر گرفته شوند. زمانی که یک صفحه وب در اختیار مدل‌های رده‌بند قرار می‌گیرد برای تعیین موضوع محتویات باید ویژگی‌های لازم از آن‌ها استخراج و در اختیار مدل قرار بگیرد. اکتساب مقدار هر یک از ویژگی‌های یک نمونه، هزینه‌ای به همراه دارد که آن را «هزینه آزمون» می‌نامند. روش‌هایی که برای رده‌بندی متون استفاده می‌شوند معمولاً از جنبه هزینه آزمون و محاسبات کم هزینه‌تر از روش‌های رده‌بندی تصویر هستند. در این مقاله نیز جهت رسیدن به کارایی مطلوب از روش‌های پرهزینه‌تر برای رده‌بندی تصاویر استفاده شده است که هزینه آزمون و محاسبات در آن قابل توجه است. هزینه دیگر در این کارکرد، هزینه رده‌بندی نادرست است. با فرض داشتن $\boldsymbol{n}$ رده‌ی متفاوت، در حالت کلی یک ماتریس $\boldsymbol{n \times n}$ خواهیم داشت که عنصر موجود در سطر $\boldsymbol{i}$ و ستون $\boldsymbol{j}$ این ماتریس تعیین کننده‌ی هزینه‌ی انتساب اشتباه یک نمونه به رده‌ی $\boldsymbol{i}$، با رده‌ی واقعی $\boldsymbol{j}$ است. برای فرآیند فرماندهی و کنترل پایش وب $\boldsymbol{W}$ خواهیم داشت:

$$\begin{aligned}\boldsymbol{Cost_{bs\_monitoring}(W)} &= \boldsymbol{Cost_{missclassification\_errors}(W)} \\ &+ \boldsymbol{Cost_{tests\_ontest}(W)} \\ &+ \boldsymbol{Cost_{computations\_ontest}(W)}\end{aligned} \quad (٥)$$

که در آن $\boldsymbol{Cost_{missclassification\_errors}}$ هزینه خطای رده‌بندی نادرست، $\boldsymbol{Cost_{tests\_ontest}}$ هزینه آزمون یا استخراج ویژگی‌ها در زمان آزمایش، و $\boldsymbol{Cost_{computations\_ontest}}$ هزینه‌های محاسبات در زمان آزمایش است.

۳-۵-۳. هزینه درک موقعیت

در این کارکرد، احتمال مربوط بودن صفحه جاری به موضوع هدف و احتمال رسیدن به صفحات هدف با دنبال نمودن پیوندهای جاری تعیین می‌شود. هزینه مربوط به تعیین این عدم قطعیت‌ها در هزینه خطای رگرسیون که در واقع خطای احتمال‌های محاسبه شده را نشان می‌دهد، انعکاس می‌یابد. برای فرآیند فرماندهی و کنترل پایش وب $\boldsymbol{W}$ داریم:

$$\boldsymbol{Cost_{understanding\_ms}(W) = Cost_{regression\_errors}(W)} \quad (٦)$$

که در آن $\boldsymbol{Cost_{regression\_errors}}$ هزینه خطاهای رگرسیون است.

۳-۵-٤. هزینه حس‌سازی

در کارکرد حس‌سازی پایش وب، هزینه اصلی مربوط به انجام فرآیند ارزیابی پیوندهای دنبال شده توسط عامل‌های پایشگر است. این کار می‌تواند توسط عامل انسانی یا به‌صورت خودکار انجام پذیرد. استفاده از عامل انسانی که شامل هزینه **HCI** است نتایج قطعی‌تری ایجاد می‌کند اما برای حجم زیاد صفحات پایش شده این هزینه بسیار زیاد خواهد بود، و برای عملیاتی بودن سامانه، جز در موارد خاص باید از عامل‌های هوشمند با ماهیت سایبری برای کنترل فرآیند پایش استفاده شود. از آنجا که حین انجام کارکرد آگاهی، رده‌ی مربوط به محتویات صفحات جاری توسط مدل‌های آموزش دیده تعیین می‌شود، این رده‌بندی می‌تواند به‌عنوان معیار ارزیابی پیوند دنبال شده توسط عامل‌های پایشگر مورد استفاده قرار بگیرد. به این ترتیب بدون صرف هزینه جدید در این کارکرد فرآیند ارزیابی تصمیمات

عامل‌های پایشگر انجام می‌شود. برای فرآیند فرماندهی و کنترل پایش وب $\boldsymbol{W}$ خواهیم داشت:

$$\boldsymbol{Cost_{sensmaking}(W)} = \boldsymbol{Cost_{hci\_evaluation}(W)} + \boldsymbol{Cost_{computations\_evaluation}(W)} \quad (۷)$$

که در آن $\boldsymbol{Cost_{hci\_evaluation}}$ هزینه‌ی ارزیابی با استفاده از **HCI** و $\boldsymbol{Cost_{computations\_evaluation}}$ هزینه ارزیابی پیش‌بینی انجام شده در مورد محتویات صفحات است.

۳-۵-۵. هزینه توسعه قصد فرمان

هزینه‌های اصلی در این کارکرد مربوط به فرآیند بازنمایی قصد فرمان در قالب مدل‌های آموزش دیده برای پایش وب است که شامل هزینه نمونه‌ها، مربی، و هزینه آزمون‌ها و محاسبات در زمان آموزش است. در مسائل دنیای واقعی دست‌یابی به نمونه‌های آموزشی، هزینه در بر دارد. در مسئله پایش وب، بخشی از نمونه‌های آموزشی حین کارکرد دیدبانی فضای نبرد تأمین می‌شوند و هزینه تأمین آن‌ها با هزینه این کارکرد همپوشانی دارد؛ اما در صورت نیاز به دریافت نمونه‌های آموزشی بیشتر از فضای وب هزینه پهنای باند مربوط به دریافت این نمونه‌ها محاسبه می‌شود. با صرف هزینه برای مربی که یک عامل انسانی است، برای نمونه‌های آموزشی برچسب تعیین می‌شود. این کار برای آموزش مدلی قابل اتکا جهت شناسایی محتویات هدف لازم است. اگر نمونه‌ها از منابعی دریافت شوند که موضوعات صفحات را از پیش توسط کاربر انسانی تعیین کرده‌اند، هزینه مربی از فرآیند حذف می‌شود.

در فرآیند آموزش مدل‌ها برای شناسایی قالب‌های مختلف محتویات هدف شامل متن و تصویر نیز هزینه‌هایی برای انجام آزمون‌ها یا همان استخراج ویژگی‌ها، و انجام محاسبات وجود دارد. در روش پیاده‌سازی شده این مقاله، هزینه‌های مذکور برای مدل‌های رده‌بند متن ناچیز ولی برای مدل‌های رده‌بند تصویر قابل توجه است و مورد بررسی بیشتر قرار گرفته است. همچنین در این کارکرد برای تعیین موضوعات هدف عامل‌های پایشگر نیاز به صرف هزینه محدودی از نوع **HCI** است. برای فرآیند فرماندهی و کنترل پایش وب $\boldsymbol{W}$ خواهیم داشت:

$$\boldsymbol{Cost_{command\_intent}(W)} = \boldsymbol{Cost_{cases}(W)} + \boldsymbol{Cost_{teacher}(W)} + \boldsymbol{Cost_{tests\_ontrain}(W)} + \boldsymbol{Cost_{computations\_ontrain}(W)} \quad (۸)$$

که در آن $\boldsymbol{Cost_{cases}}$ هزینه نمونه‌ها، $\boldsymbol{Cost_{teacher}}$ هزینه مربی، $\boldsymbol{Cost_{tests\_ontrain}}$ هزینه آزمون یا استخراج ویژگی‌ها در زمان آموزش، و $\boldsymbol{Cost_{computations\_ontrain}}$ هزینه‌های محاسبات در زمان آموزش است.

۳-۵-۶. هزینه مدیریت فضای نبرد

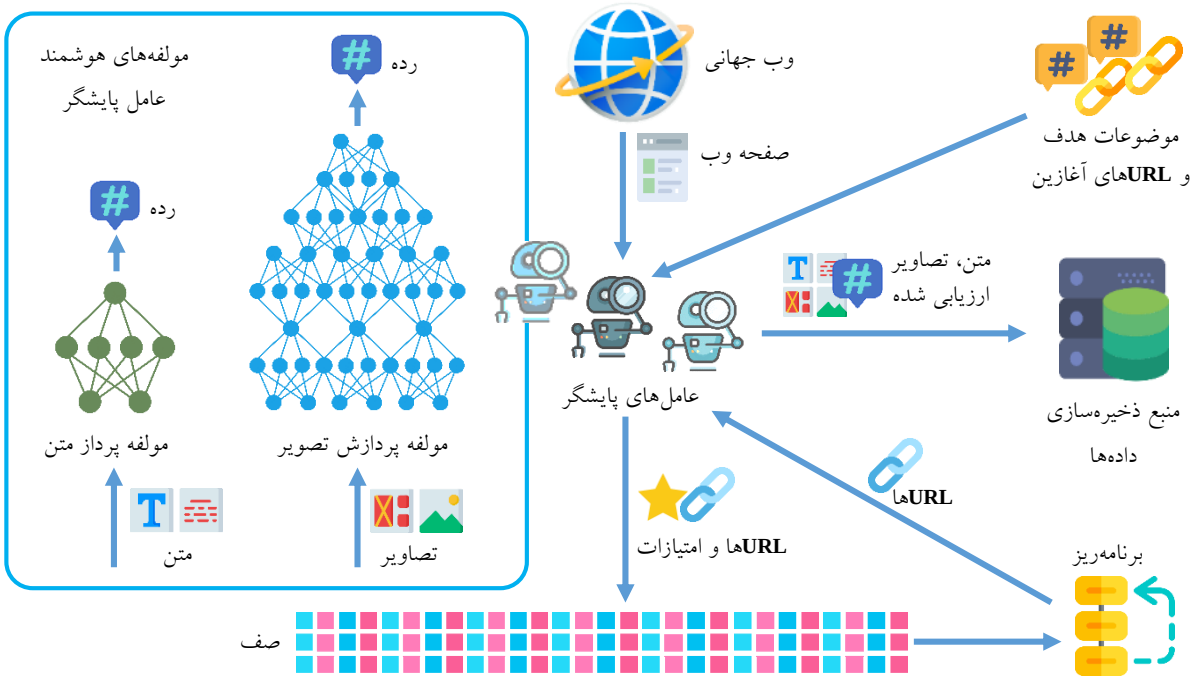
بخش اصلی هزینه‌ها در این کارکرد شامل هزینه‌های محاسباتی زیرساخت مدیریت و زمان‌بندی کارهایی است که باید توسط عامل‌های پایشگر انجام شود. در مسئله پایش وب بررسی شده در این مقاله، هزینه انجام این محاسبات به میزان قابل ملاحظه‌ای کمتر از هزینه محاسبات لازم برای آموزش و به‌کارگیری مدل‌های رده‌بند است. برای فرآیند فرماندهی و کنترل پایش وب $\boldsymbol{W}$ خواهیم داشت:

$$\boldsymbol{Cost_{bs\_management}(W)} = \boldsymbol{Cost_{computation\_scheduling}(W)} \quad (۹)$$

که در آن $\boldsymbol{Cost_{computation\_scheduling}}$ هزینه محاسبات برای برنامه‌ریزی عامل‌های پایشگر است.

٤. سامانه C4ISR پیاده‌سازی شده پایش وب

برمبنای مفاهیم نظری تشریح شده در بخش قبل، یک سامانه‌ی **C4ISR** پایش وب پیاده‌سازی شده است که در این قسمت به معرفی و بررسی اجزای آن پرداخته می‌شود. این سامانه بر مبنای یک رویکرد چندعامله و حساس به هزینه برای فرماندهی و کنترل پایش وب توسعه داده شده است. با در نظر گرفتن بحث هزینه در سامانه مذکور و جهت دستیابی به سطح مطلوبی از کارایی، روش‌های مورد استفاده در دو مؤلفه هزینه‌بر سامانه یعنی پردازش تصویر و خزش موضوعی، مورد بررسی بیشتر قرار گرفته‌اند و برای هر دو جزء توسط این مقاله روشی پیشنهاد شده است که با رویکرد حساس به هزینه، بیشترین میزان کارایی را در اختیار سامانه قرار می‌دهد. در واقع هدف اصلی در این بخش دست‌یابی به یک سامانه **C4ISR** پایش وب با کمترین هزینه و بیشترین کارایی بر مبنای معیارهای ارزیابی استاندارد است. در ادامه ابتدا روش‌های استفاده شده در سامانه را مرور و



شکل ۳. ساختار سامانه **C4ISR** پیاده‌سازی شده برای پایش محتوای وب.

ساختار آن را ارائه می‌کنیم، سپس روش پیشنهادی برای مؤلفه پردازش تصویر و خزش موضوعی سامانه را توضیح می‌دهیم.

٤-١. ساختار و فرآیند سامانه C4ISR پایش وب

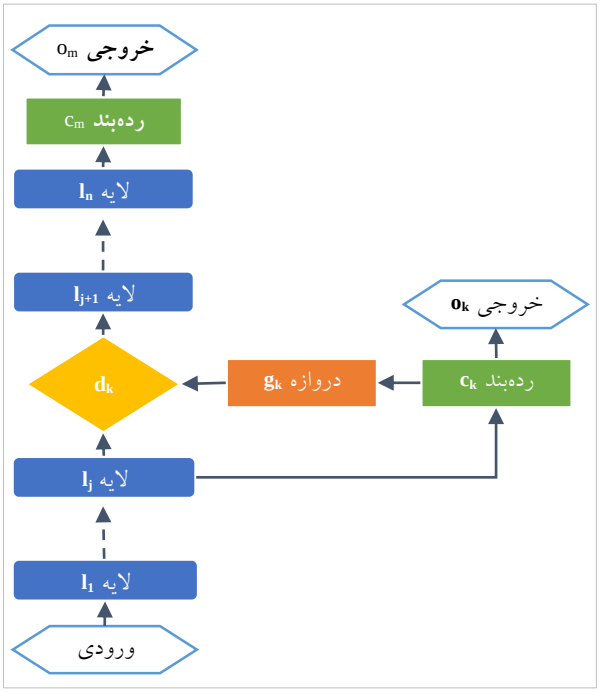
شکل ۳ نمایی از ساختار سامانه **C4ISR** پیاده‌سازی شده برای پایش محتوای وب را نشان می‌دهد. ابتدا مجموعه‌ای از موضوعات هدف و **URL**های آغازین، با استفاده از منابع موجود روی وب به کمک عامل انسانی تعیین می‌شود. قصد فرمان با مدل‌سازی موضوعات مذکور با استفاده از روش‌های یادگیری رده‌بند، بیان می‌شود و در مؤلفه‌های هوشمند عامل‌های پایشگر یعنی مدل پردازش متن و مدل پردازش تصویر انعکاس می‌یابد. عامل‌های پایشگر بر مبنای موارد مذکور، فرآیند خزش وب و دیدبانی فضای نبرد را آغاز می‌کنند. با دریافت هر صفحه، محتوای آن توسط مؤلفه‌های هوشمند یا همان رده‌بندهای تحلیلگر بررسی می‌شود و با انجام رده‌بندی، فرآیند آگاهی صورت می‌گیرد. سپس طی فرآیند درک موقعیت، میزان قطعیت نتایج رده‌بندی با روش رگرسیون برای محتویات صفحه تعیین می‌شود و امتیاز مربوط به دنبال نمودن هر یک از پیوندهای موجود در صفحه پیش‌بینی می‌گردد. در فرآیند حس‌سازی، میزان مربوط بودن صفحه دریافت شده به موضوع هدف با مقدار امتیازی که پیش‌تر برای پیوند متناظر با آن پیش‌بینی شده است مقایسه می‌گردد و انتخاب مذکور مورد ارزیابی قرار می‌گیرد. همچنین پیوندهای مناسب موجود در صفحه برای رسیدن به صفحات هدف در فضای وب انتخاب می‌شوند. محتویات ارزیابی شده جهت استفاده و انجام پردازش‌های بعدی در منبع ذخیره‌سازی داده‌ها جای می‌گیرند. مؤلفه برنامه‌ریز سامانه، وظیفه مدیریت صف پیوندها و زمان‌بندی پردازش‌های محاسباتی جهت انجام فرآیند توسط عامل‌های پایشگر را در قالب کارکرد مدیریت فضای نبرد بر عهده دارد.

٤-٢. روش پیشنهادی پردازش تصویر حساس به هزینه

همان طور که گفته شد تصاویر موجود در صفحات وب از عناصر اصلی تعیین کننده موضوع صفحه محسوب می‌شوند و پرهزینه‌ترین بخش هزینه آزمون و محاسبات یک مدل را به خود اختصاص می‌دهند. **CNN**های عمیق در حال حاضر کارآمدترین روش برای رده‌بندی تصاویر محسوب می‌شوند. روش پیشنهادی نیز یک **CNN** عمیق حساس به هزینه است. لازم به ذکر است روش پیشنهادی قادر است با اغلب روش‌های توصیف شده در بخش‌های قبل ترکیب شود و مدل‌هایی با هزینه پردازش کمتر ایجاد نماید.

٤-٢-١. ساختار روش پیشنهادی پردازش تصویر

شکل ٤ نمایی از ساختار **CNN** عمیق حساس به هزینه را نشان می‌دهد.



شکل ٤. ساختار پیشنهادی **CNN** عمیق حساس به هزینه

با در نظر گرفتن شکل می‌توان مجموعه عناصر شبکه را به‌صورت زیر تعریف کرد:

- مجموعه‌ی $\boldsymbol{L}$ شامل $\boldsymbol{n}$ لایه‌ی معمول شبکه‌های **CNN**: $\boldsymbol{L = \{l_1, l_2, \dots, l_n\}}$.
- مجموعه $\boldsymbol{C}$ شامل $\boldsymbol{m}$ رده‌بند: $\boldsymbol{C = \{c_1, c_2, \dots, c_m\}}$. همه‌ی رده‌بندها به‌جز $\boldsymbol{c_m}$، از نوع رده‌بندهای میانی شبکه هستند.
- مجموعه $\boldsymbol{G}$ شامل $\boldsymbol{m}$ دروازه[1]: $\boldsymbol{G = \{g_1, g_2, \dots, g_m\}}$. دروازه‌ها در واقع توابعی هستند که در مورد ادامه فرآیند استخراج ویژگی‌ها و انجام محاسبات در لایه‌های بالاتر از خود تصمیم می‌گیرند.
- مجموعه خروجی‌های $\boldsymbol{O}$: $\boldsymbol{O = \{o_1, o_2, \dots, o_m\}}$ به‌طوری که هر خروجی $\boldsymbol{o_i}$ توسط یک رده‌بند $\boldsymbol{c_i}$ محاسبه می‌شود.
- مجموعه تصمیم‌های $\boldsymbol{D}$: $\boldsymbol{D = \{d_1, d_2, \dots, d_m\}}$ به‌طوری که هر تصمیم $\boldsymbol{d_i}$ توسط یک دروازه $\boldsymbol{g_i}$ محاسبه می‌شود. در واقع مقادیر این تصمیمات تعیین کننده ادامه و یا توقف فرآیند پردازش در لایه‌های بالاتر شبکه است.

٤-٢-٢. الگوریتم روش پیشنهادی پردازش تصویر

فرآیند رده‌بندی یک ورودی در مدل پیشنهادی **CNN** عمیق حساس به هزینه در قالب یک الگوریتم در شکل ٥ نمایش داده شده است. طبق مراحل ارائه شده برای نحوه عملکرد **CNN** عمیق حساس به هزینه، پس از دریافت نمونه ورودی، ابتدا فرآیند محاسبه مقادیر لایه‌ها تا اولین لایه متصل به یک رده‌بند میانی شبکه انجام می‌شود. سپس خروجی دروازه جاری محاسبه می‌گردد. اگر تصمیم دروازه توقف فرآیند باشد، خروجی رده‌بند جاری محاسبه و به‌عنوان خروجی نهایی ارائه می‌شود. در غیر این صورت فرآیند محاسبه مقادیر لایه‌ها تا اولین لایه بعدی متصل به یک رده‌بند میانی شبکه انجام می‌شود. اگر به رده‌بند نهایی برسیم، فرآیند پایان خواهد یافت.

١. دریافت یک نمونه در ورودی شبکه.
٢. قرار بده:
   - $\boldsymbol{i \leftarrow 1}$
   - $\boldsymbol{k \leftarrow 1}$
   - $\boldsymbol{j \leftarrow}$ "اولین لایه متصل به $\boldsymbol{c_k}$"
٣. پردازش ورودی و محاسبه مقادیر لایه‌های شبکه از لایه $\boldsymbol{l_i}$ تا $\boldsymbol{l_j}$ که خروجی آن متصل به پیمانه‌ی رده‌بند $\boldsymbol{c_k}$ است.
٤. اگر $\boldsymbol{k == m}$:
   - محاسبه خروجی $\boldsymbol{c_k}$ یعنی $\boldsymbol{o_k}$.
   - خروجی نهایی را برابر $\boldsymbol{o_k}$ قرار بده. **پایان**.
٥. محاسبه خروجی $\boldsymbol{g_k}$ یعنی $\boldsymbol{d_k}$.
٦. اگر $\boldsymbol{d_k}$ ادامه فرآیند استنتاج است:
   - $\boldsymbol{i \leftarrow j}$
   - $\boldsymbol{k \leftarrow k + 1}$
   - $\boldsymbol{j \leftarrow}$ "اولین لایه متصل به $\boldsymbol{c_k}$"
   - برو به ٣.
٧. اگر $\boldsymbol{d_k}$ ادامه فرآیند استنتاج نیست:
   - محاسبه خروجی $\boldsymbol{c_k}$ یعنی $\boldsymbol{o_k}$.
   - خروجی نهایی را برابر $\boldsymbol{o_k}$ قرار بده. **پایان**.

شکل ٥. الگوریتم رده‌بندی در مدل پیشنهادی **CNN** عمیق حساس به هزینه

به این ترتیب با مدیریت هزینه محاسبات و استخراج مقادیر ویژگی‌ها، شبکه قادر خواهد بود نمونه‌های آسان را با صرف

[1] Gate

هزینه کمتری رده‌بندی نماید، و برای نمونه‌های پیچیده‌تر هزینه بیشتری صرف کند و انجام محاسبات را در لایه‌های بالاتر شبکه ادامه دهد.

٤-٣. روش پیشنهادی برای خزش وب حساس به هزینه

در قسمت‌های قبل روش‌های مختلف استخراج زمینه پیوند برای خزش موضوعی وب تشریح شد. به‌طور خلاصه می‌توان گفت چالش روش متن صفحه، یکسان در نظر گرفتن زمینه پیوند یعنی متن کل صفحه برای همه‌ی پیوندها است. چالش روش پنجره متن، نامعلوم بودن اندازه بهینه پنجره و محل آن نسبت به پیوند است. چالش روش مبتنی بر بلوک **VIPS** نیز استخراج بلوک‌های نادرست بیش از حد بزرگ برای برخی از پیوندها است که موجب ایجاد خطا در استخراج زمینه پیوند و افزایش هزینه خزش وب می‌گردد. در این مقاله برای حل مشکلات روش‌های مذکور که از بهترین روش‌های خزش موضوعی محسوب می‌شوند و بالاترین سطح نتایج را به خود اختصاص داده‌اند [۱۸]، یک روش ترکیبی ساده اما کارآمد برای کاهش خطای کلی خزشگرها پیشنهاد شده است و با ارزیابی عملی روش پیشنهادی، برتری آن نسبت به سایر روش‌های موجود نمایش داده شده است. روش ترکیبی پیشنهادی برای محاسبه امتیاز یک پیوند بر اساس رابطه زیر عمل می‌کند:

$$\boldsymbol{link\_score} = \frac{\sum_{\boldsymbol{lc}\in\{\boldsymbol{link\_contexts}\}} \boldsymbol{Relevancy(lc)}}{|\boldsymbol{link\_contexts}|} \qquad (۱۰)$$

و داریم:

$$\begin{aligned}\sum_{\boldsymbol{lc}\in\{\boldsymbol{link\_contexts}\}} \boldsymbol{Relevancy(lc)} &= \boldsymbol{Relevancy(page\_text)} \\ &+ \big(\boldsymbol{Relevancy(link\_text)} \\ &+ \boldsymbol{Relevancy(window10\_text)} \\ &+ \boldsymbol{Relevancy(window20\_text)} \\ &+ \boldsymbol{Relevancy(window40\_text)} \\ &+ \boldsymbol{Relevancy(vips\_block\_text)}\big)\end{aligned} \qquad (۱۱)$$

که در آن $\boldsymbol{lc}$ زمینه پیوند، $\{\boldsymbol{link\_contexts}\}$ مجموعه‌ی زمینه پیوندهای استخراج شده با استفاده از روش‌های مختلف، و $|\boldsymbol{link\_contexts}|$ نشان‌دهنده تعداد روش‌های مذکور است. همچنین $\boldsymbol{window10\_text}$، $\boldsymbol{window20\_text}$ و $\boldsymbol{window40\_text}$، متن استخراج شده به‌عنوان زمینه پیوند با استفاده از روش پنجره متن به ترتیب برای پنجره‌هایی با اندازه ۱۰، ۲۰ و ٤٠ است. $\boldsymbol{vips\_block\_text}$ نیز زمینه پیوند استخراج شده با روش مبتنی بر بلوک الگوریتم **VIPS** را نشان می‌دهد. در واقع رابطه ۱۰ میانگین امتیاز به دست آمده برای زمینه پیوندهای استخراج شده با روش‌های مختلف را محاسبه می‌کند.

مبنای ایده‌ی این نحوه‌ی ترکیب امتیاز حاصل از روش‌های استخراج زمینه پیوند، رابطه‌ی ۱ و نتایج گزارش شده برای مقدار مناسب ضریب $\boldsymbol{\beta}$ است که برابر با ۰/۲۵ در نظر گرفته شده است [۱۷]. طبق نتایج مذکور زمانی که امتیاز متن صفحه با متن پیوند به نحوی ترکیب شود که متن پیوند وزن بیشتری را به خود اختصاص دهد و تأثیر بیشتری در امتیاز نهایی داشته باشد، کارایی خزشگر موضوعی نسبت به استفاده تنها از هر یک از این دو روش استخراج زمینه پیوند بالاتر خواهد بود. بر این اساس در روش پیشنهادی مقاله حاضر به‌جای وزن‌دهی یک‌باره و بیشتر به متن پیوند در رابطه ۱، امتیاز متن پیوند، سه پنجره متن با اندازه‌های مختلف، و متن بلوک احاطه کننده پیوند با استفاده از میانگین‌گیری باهم ترکیب شده است. از آنجا که روش‌های مذکور در پژوهش‌های پیشین در مقایسه با متن پیوند عملکرد بهتری از خود نشان داده‌اند و می‌توانند با دقت بهتری زمینه پیوند را شناسایی کنند، انتظار می‌رود میانگین آن‌ها نیز در ترکیب با یکدیگر و امتیاز متن صفحه، با پوشش ضعف‌های منفرد هر یک از این روش‌ها، به کاهش هزینه خزش موضوعی وب منجر گردد؛ همان‌گونه که ترکیب امتیاز متن صفحه و متن پیوند نقاط ضعف استفاده تنها از این دو روش را پوشش داده است. نتایج ارائه شده در این مقاله برتری روش پیشنهادی و ایده مطرح شده را نسب به سایر روش‌ها تأیید می‌کند.

۵. نتایج پیاده‌سازی سامانه پیشنهادی

در این بخش، شرایط پیاده‌سازی و نتایج ارزیابی بخش‌های پردازش متن، پردازش تصویر، و خزشگر موضوعی سامانه پیشنهادی **C4ISR** پایش وب بررسی می‌شود. با توجه به گستردگی بحث ارزیابی هزینه در بخش‌های مختلف سامانه، در

این مقاله هزینه‌های عموماً مؤثرتر در فرآیند فرماندهی و کنترل برای کارکردهای زیر ارزیابی شده است:

- کارکرد دیدبانی فضای نبرد: هزینه‌ی پهنای باند در قالب دو معیار نرخ درو[1] و بازخوانی هدف[2] ارزیابی شده است.
- کارکرد آگاهی: خطای رده‌بندی نادرست با در نظر گرفتن هزینه خطای یکسان برای انتساب نادرست رده‌ها، با استفاده از معیارهای بازخوانی[3]، دقت[4]، معیار **F**[5]، و صحت[6] که هر یک به‌نوعی معکوس خطا را نشان می‌دهند، ارزیابی شده است. با توجه به ناچیز بودن هزینه‌ی آزمون‌ها و محاسبات زمان آزمایش برای مؤلفه پردازش متن، این هزینه‌ها برای مؤلفه پردازش تصویر در قالب زمان پردازش مورد بررسی قرار گرفته است.
- کارکرد حس‌سازی: با توجه به استفاده از نتایج رده‌بندهای به کار رفته در کارکرد آگاهی، برای انجام فرآیند ارزیابی پیش‌بینی‌ها، و عدم استفاده از عامل انسانی برای انجام این کار، هزینه این کارکرد در پیاده‌سازی انجام شده معادل با هزینه آگاهی است.
- کارکرد توسعه قصد فرمان: هزینه نمونه‌ها و مربی، به دلیل دریافت اطلاعات نمونه‌های برچسب‌گذاری شده از **DMOZ**، برای پیاده‌سازی انجام شده برابر با صفر است.

همان‌طور که توضیح داده شد هزینه‌های اصلی که مربوط به مؤلفه‌های پردازش متن، پردازش تصویر و خزش موضوعی سامانه هستند مورد ارزیابی دقیق قرار گرفته است.

۵-۱. معیارهای ارزیابی

بر مبنای رویکرد حساس به هزینه پژوهش حاضر که در قسمت‌های قبل توضیح داده شد، علاوه بر ارزیابی کارایی سامانه با معیارهای استاندارد متداول، بحث هزینه‌های محاسباتی و نحوه‌ی استفاده از منابع نیز مورد بررسی قرار می‌گیرد. برای ارزیابی میزان درستی بخش پردازش متن و پردازش تصویر سامانه از معیارهای شناخته شده و استاندارد بازخوانی، دقت، و معیار **F** استفاده شده است. با توجه به کم هزینه بودن پردازش‌های استفاده شده برای رده‌بندهای پردازش متن و قابل اغماض بودن آن در مقابل هزینه‌های مربوط به پردازش تصویر، هزینه مربوط به انجام محاسبات سامانه برای بخش پردازش تصاویر بر اساس معیار زمان مورد بررسی قرار گرفته است. همچنین نتایج این بخش با معیار صحت نیز مورد سنجش قرار گرفته است. معیارهای مذکور با استفاده از روابط زیر محاسبه می‌شود.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}} \qquad (۱۲)$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}} \qquad (۱۳)$$

$$\boldsymbol{F\_measure} = \frac{\boldsymbol{2}}{\boldsymbol{Recall^{-1} + Precision^{-1}}} \qquad (۱۴)$$

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \qquad (۱۵)$$

که در این روابط، $\boldsymbol{Recall}$ بازخوانی، $\boldsymbol{Precision}$ دقت، $\boldsymbol{F\_measure}$ معیار **F**، و $\boldsymbol{Accuracy}$ صحت است. همچنین $\boldsymbol{TP}$ مثبت درست، $\boldsymbol{TN}$ منفی درست، $\boldsymbol{FP}$ مثبت نادرست، و $\boldsymbol{FN}$ منفی نادرست است.

از آنجا که هزینه اصلی در بخش خزشگر سامانه مربوط به نحوه استفاده از پهنای باند اینترنت برای پایش وب است، برای ارزیابی خزشگر از دو معیار استاندارد استفاده می‌کنیم که میزان کارایی این بخش را بر اساس حداقل اتلاف در منابع پهنای باند تعریف می‌کنند و هم زمان دقت فرآیند خزش را نشان می‌دهند. این دو معیار، نرخ درو و بازخوانی هدف نام دارند و پژوهش‌های مختلف از آن‌ها برای ارزیابی خزشگرهای موضوعی استفاده نموده‌اند [۱۷].

نرخ درو برای $\boldsymbol{t}$ صفحه‌ی دریافت شده از ابتدا تا کنون یعنی $\boldsymbol{Harvest\_Rate(t)}$ با استفاده از رابطه‌ی زیر محاسبه می‌شود:

$$\boldsymbol{Harvest\_Rate(t)} = \frac{\boldsymbol{1}}{\boldsymbol{t}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{t}} \boldsymbol{r_i} \qquad (۱۶)$$

که در آن $\boldsymbol{r_i}$ امتیاز مربوط بودن صفحه‌ی دریافتی $\boldsymbol{i}$ام بر اساس خروجی رده‌بند ارزیاب به شکل دودویی (صفر و یک) است.

[1] Harvest Rate
[2] Target Recall
[3] Recall
[4] Precision
[5] F-measure
[6] Accuracy

معیار بازخوانی هدف برای ارزیابی میزان کارایی خزشگرهای موضوعی، ابتدا مجموعه‌ای از صفحات هدف مربوط به موضوع، یعنی $\boldsymbol{T}$ را تعیین می‌کند. سپس در هر زمان عملکرد خزشگر را بر اساس نسبت تعداد صفحات خزیده شده‌ی موجود در $\boldsymbol{T}$ به تعداد کل صفحات موجود در $\boldsymbol{T}$، می‌سنجد. به بیان دیگر اگر $\boldsymbol{Target\_Recall(t)}$، بازخوانی هدف خزنده‌های متمرکز برای $\boldsymbol{t}$ صفحه‌ی دریافت شده تا کنون باشد داریم:

$$\boldsymbol{Target\_Recall(t)} = \frac{|\boldsymbol{C(t)} \cap \boldsymbol{T}|}{|\boldsymbol{T}|} \qquad (۱۷)$$

که در آن $\boldsymbol{C(t)}$ مجموعه صفحات خزیده شده از ابتدا تا صفحه‌ی $\boldsymbol{t}$ام و $|\boldsymbol{T}|$ تعداد صفحات موجود در $\boldsymbol{T}$ است.

۵-۲. موضوعات مورد بررسی و مجموعه داده‌ها

برای بخش رده‌بند متن و خزشگر موضوعی حساس به هزینه سامانه، هشت موضوع از **DMOZ** [۴۱] انتخاب شده است: الگوریتم، فوتبال انگلیس، کمک‌های اولیه، نظریه‌ی گراف، جاوا، لینوکس، المپیک، و روباتیک. صفحات مربوط به هر از این موضوعات به همراه صفحات مربوط به یک عمق مشخص از زیرموضوع‌های آن‌ها بر اساس ساختار سلسله مراتبی **DMOZ** از وب استخراج شده‌اند. صفحات استخراج شده، مجموعه صفحات مربوط به هر یک از موضوعات را شکل می‌دهند. صفحات مربوط به خود موضوع، ریشه با عمق صفرِ موضوع را بر روی ساختار سلسله مراتبی **DMOZ** تشکیل می‌دهند، صفحات مربوط به خود موضوع به همراه زیرموضوعات مستقیم آن، عمق یکِ موضوع را تشکیل می‌دهند و همین طور برای سایر زیرموضوع‌ها. مشابه این روش در [۴۲] برای استخراج بردار بازنمایی موضوع از توضیحات ارائه شده برای صفحات مرتبط به موضوع در **DMOZ** پیشنهاد شده است.

ویژگی‌های موضوعات بررسی شده در جدول ۲ نشان داده شده است. تعداد صفحات، نشان‌دهنده‌ی تعداد صفحات مربوط به موضوع از ریشه تا عمق تعیین شده بر اساس ساختار **DMOZ** است. آدرس صفحات از [۴۱] استخراج شده است.

برای محاسبه‌ی معیار بازخوانی هدف، کسری از مجموعه صفحات مربوط به هر موضوع به‌عنوان صفحات هدف در نظر گرفته شده است. تعداد این صفحات به همراه **URL**های آغازین هر یک از موضوعات که صف اولویت پیوندها با آن‌ها مقداردهی اولیه می‌شود، در جدول ۳ آمده است.

جدول ۲. توصیف موضوعات انتخاب شده از DMOZ

موضوع	عمق‌در **DMOZ**	تعداد صفحات	**URL** ریشه در **DMOZ**
الگوریتم	۱	۱۲۰	**Computers/Algorithms**
فوتبال انگلیس	۱	۲۵۰	**Sports/Soccer/UEFA/England**
کمک‌های اولیه	۱	۴۰	**Health /Public_Health_and_Safety /First_Aid**
نظریه گراف	۲	۸۰	**Science/Math/Combinatorics /Graph_Theory**
جاوا	۱	۴۱۰	**Computers/Programming /Languages/Java**
لینوکس	۱	۳۶۰	**Computers/Software /Operating_Systems/Linux**
المپیک	۳	۳۳۰	**Sports/Events/Olympics**
روباتیک	۱	۳۵۰	**Computers/Robotics**

جدول ۳. URLهای آغازین و تعداد صفحات هدف برای موضوعات

موضوع	صفحات هدف	**URL**های آغازین
الگوریتم	۴۰	**www.ask.com/web?q=algorithms www.google.com/search?q=algorithms en.wikipedia.org/wiki/Algorithm**
فوتبال انگلیس	۸۰	**dir.yahoo.com/Regional/Countries /United_Kingdom/Recreation_and_Sports /Sports/Soccer/ (from archive.org) www.ask.com/web?q=soccer+england**
کمک‌های اولیه	۱۵	**dir.yahoo.com/Health/First_Aid/ www.ask.com/web?q=first+aid**
نظریه گراف	۳۰	**www.ask.com/web?q=graph+theory www.google.com/search?q=graph+theory en.wikipedia.org/wiki/Graph_theory**
جاوا	۱۴۰	**dir.yahoo.com/Computers_and_Internet /Programming_and_Development /Languages/Java/ (from archive.org) www.ask.com/web?q=java+programming**
لینوکس	۱۲۰	**dir.yahoo.com/Computers_and_Internet /Software/Operating_Systems/UNIX/Linux (from archive.org) www.ask.com/web?q=linux**
المپیک	۱۱۰	**dir.yahoo.com/Recreation/Sports/Events /International_Games/Olympic_Games/ (from archive.org) www.ask.com/web?q=olympic+games**
روباتیک	۱۲۰	**dir.yahoo.com/Science/Engineering /Mechanical_Engineering/Robotics/ (from archive.org) http://www.ask.com/web?q=robotics**

برای ارزیابی بهتر بخش پردازش تصویر سامانه، مجموعه داده‌ای با بیش از ۱۰۰ هزار تصویر شامل تصاویر با محتویات عادی و تصاویر حاوی خشونت توسط کاربر انسانی جمع‌آوری و برچسب‌گذاری شده است. حدود ۸۵٪ از این داده‌ها که شامل طیف متنوعی از تصاویر است و از منابع مختلف جمع‌آوری شده است، برای آموزش و مابقی برای ارزیابی روش‌های پردازش تصویر مورد استفاده قرار گرفته است. جدول ٤ مشخصات مجموعه داده مذکور را نشان می‌دهد.

جدول ٤. مشخصات مجموعه داده‌ها برای بخش پردازش تصویر

	مجموعه داده آموزش	مجموعه داده آزمایش
تعداد تصاویر عادی	٦۰۰۰۰	۱۰۰۰۰
تعداد تصاویر خشن	۳۰۰۰۰	۵۰۰۰
جمع کل	۹۰۰۰۰	۱۵۰۰۰

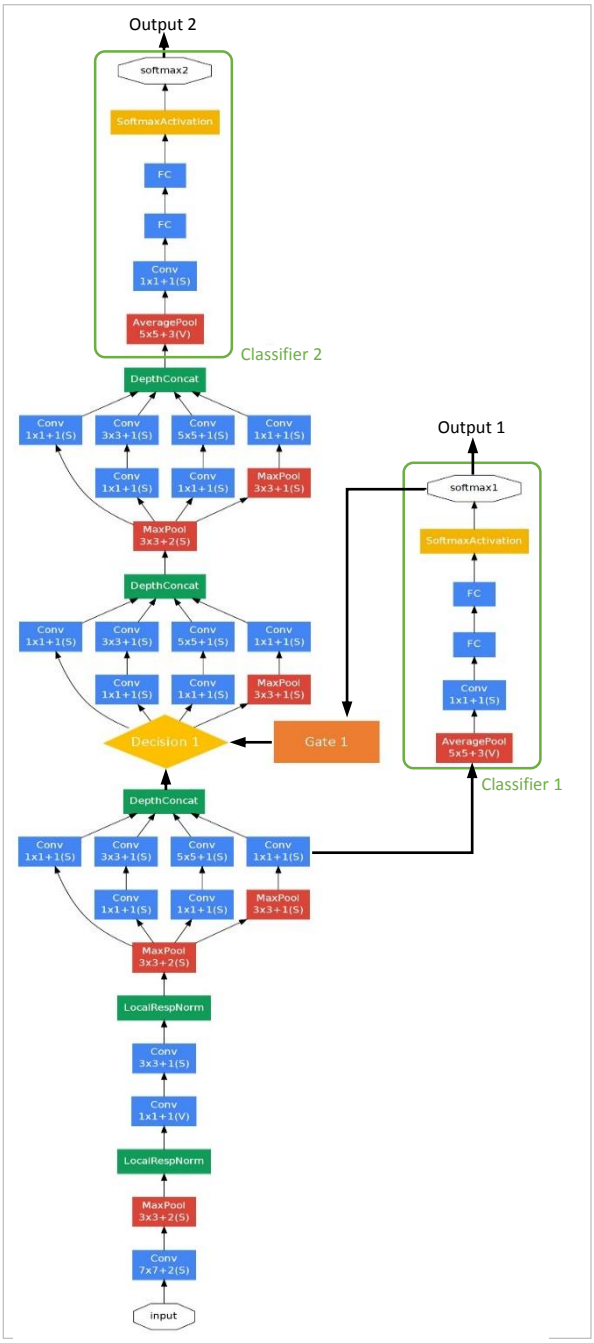
۵-۳. تنظیمات و جزئیات پیاده‌سازی

رده‌بندهای تحلیلگر و ارزیاب متن برای هر یک از موضوعات مورد بررسی، از نوع شبکه‌های عصبی پرسپترون چندلایه[1] (**MLP**) هستند. بازنمایی صفحات با استفاده از مدل فضای بردار[2] (**VSM**) [٤۳] انجام شده است. برای بازنمایی صفحات در این فضا، ابتدا متن موجود در هر یک از صفحات توسط تجزیه‌کننده‌ی **HTML** [٤٤] استخراج می‌شود. سپس واژه‌های زائد[3] از متن حذف می‌گردند و سایر واژه‌ها با الگوریتم **Porter** [٤۵] ریشه‌یابی می‌شوند. وزن واژه‌ها نیز با روش **TFIDF** [٤۳] تعیین می‌گردد. هر یک از ابعاد فضای بردار را یکی از واژه‌های ریشه‌یابی شده تشکیل می‌دهند و بردار صفحه بر اساس وزن واژه‌های متناظر با هر یک از این ابعاد شکل می‌گیرد. هر یک از ابعاد **VSM** در قالب یک ویژگی در نظر گرفته می‌شود و این ویژگی‌ها به‌عنوان ورودی به رده‌بندها ارائه می‌گردند. به دلیل زیاد بودن تعداد واژه‌ها، آموزش رده‌بندهایی از نوع **MLP** با این ابعاد فضای ویژگی عملاً امکان‌پذیر نیست. یک راه حل برای مواجهه با این مشکل، استفاده از روش‌های انتخاب ویژگی[4] است. همچنین اگر انتخاب ویژگی به شکلی مناسب انجام شود موجب افزایش کارایی رده‌بند نیز خواهد شد [٤٦]. بر اساس نتایج گزارش شده در [٤٦] از روش انتخاب ویژگی بهره‌ی اطلاعات[5] (**IG**) برای کاهش ابعاد فضای ویژگی استفاده شده است. همه‌ی رده‌بندها از نوع **MLP** سه لایه (با یک لایه‌ی مخفی) هستند. تعداد نرون‌های لایه‌ی مخفی ۲۰٪ تعداد نرون‌های لایه‌ی ورودی در نظر گرفته شده است. آموزش رده‌بندها با استفاده از همه‌ی صفحات مربوط به یک موضوع به‌عنوان نمونه‌های مثبت و تعداد دو برابر آن صفحاتی از سایر موضوعات جز زیرموضوعات هدف به‌عنوان نمونه‌های منفی، انجام شده است. پیاده‌سازی رده‌بندها با استفاده از برخی بسته‌های[6] نرم افزار متن باز **weka** [٤۷] صورت گرفته است. آموزش رده‌بندها نیز با روش نرخ آموزش تنزیلی و تنظیمات پیش‌فرض **weka** انجام شده است.

برای پیاده‌سازی خزشگرها از کدهای موجود در [٤۸] کمک گرفته شده است. عامل‌های خزشگرها به شکل همگن پیاده‌سازی شده‌اند و از مجموعه تحلیلگرهایی یکسان برای امتیازدهی به پیوندها استفاده می‌کنند. آموزش رده‌بندهای تحلیلگر، که وظیفه هدایت خزشگرها را بر عهده دارند، مشابه با رده‌بندهای ارزیاب انجام شده است. برای پیاده‌سازی عملیات بلوک‌بندی صفحات بر اساس الگوریتم **VIPS** در تحلیلگرهایی که از اطلاعات مربوط به بلوک‌ها استفاده می‌کنند، از کدهای ارائه شده در [٤۹] کمک گرفته شده است. در پیاده‌سازی این الگوریتم مقدار **PDoC**[7] بر اساس مجموعه‌ای از آزمون و خطاهای شهودی برابر با ٦ در نظر گرفته شده است. متن موجود در بلوک‌ها طی شیوه‌ای مشابه با آنچه برای صفحات وب انجام می‌شود، از بلوک‌ها استخراج گردیده و مورد پردازش قرار گرفته است. پنجره‌های متن مطابق با توضیحاتی که در بخش‌های قبل آمده است، استخراج شده‌اند. بر مبنای آنچه در [۱۷] گزارش شده، اندازه‌ی پنجره‌های متن (***W***) کوچک، متوسط و بزرگ به ترتیب برابر با ۱۰، ۲۰ و ٤۰ کلمه منظور شده است. جمعیت

[1] Multi-Layer Perceptron
[2] Vector Space Model
[3] Stop word
[4] Feature selection
[5] Information Gain
[6] Package
[7] Predefined Degree of Coherence

عامل‌های خزنده نیز برابر با ۱۰ در نظر گرفته شده است. برای روش‌های پیاده‌سازی شده به‌جز روش پیشنهادی، از ترکیبی از امتیاز محاسبه شده بر مبنای متن کل صفحه و روش مربوطه برای تعیین امتیاز نهایی استفاده شده است. سهم هر یک از این امتیازات در محاسبه‌ی امتیاز نهایی طبق آنچه در [۱۷] پیشنهاد شده، برابر با ۲۵٪ برای امتیاز حاصل از متن کل صفحه و ۷۵٪ برای امتیاز حاصل از روش مربوطه در نظر گرفته شده است.



شکل ٦. مدل **CNN** پیشنهادی پیاده‌سازی شده برای مؤلفه پردازش تصویر

شکل ٦ نمایی از مدل **CNN** پیاده‌سازی شده برای مؤلفه پردازش تصویر را نشان می‌دهد. ساختار این مدل شبیه به بخش ابتدایی ساختار مدل گوگل نت است با این تفاوت که یک رده‌بند میانی به آن افزوده شده که شامل مجموعه‌ای از لایه‌ها مشابه رده‌بند نهایی موجود در انتهای مدل است.

خروجی رده‌بندهای این مدل عددی در بازه صفر تا یک است و رده‌ی مثبت برای محتویات خشن در نظر گرفته شده است. در این پیاده‌سازی برای تعیین حداقل اطمینان جهت تصمیم‌گیری در مورد ادامه یا توقف فرآیند محاسبات بعد از لایه متصل به رده‌بند میانی، دو مقدار آستانه برای خروجی رده‌بند تعیین می‌کنیم. اگر خروجی رده‌بند بالاتر یا پایین‌تر از این دو مقدار باشد یعنی به عدد صفر یا یک نزدیک باشد به‌نوعی نشان می‌دهد که رده و برچسب یک نمونه، با اطمینان بالایی تعیین شده است. در لایه آخر شبکه از تابع **softmax** استفاده شده که با استفاده از رابطه زیر محاسبه می‌شود.

$$\sigma(x)_i = \frac{exp(x_i)}{\sum_{j=1}^{K} exp(x_c)} \quad , \quad i = 1, \dots, K \qquad (۱۸)$$

که در آن $x = (x_1, \dots, x_K) \epsilon \mathbb{R}^K$ بردار ورودی به لایه‌ی **softmax** و K تعداد رده‌های مسئله است که در اینجا برابر با ۲ است. روش **CNN** عمیق پیشنهادی با استفاده از بسته نرم‌افزاری یادگیری عمیق **Caffe** [۵۰] و بر روی **GPU** پیاده‌سازی شده است. مدل، با استفاده از یک سیستم پنتیوم ۷ با یک کارت گرافیک انویدیا تایتان آموزش دیده است.

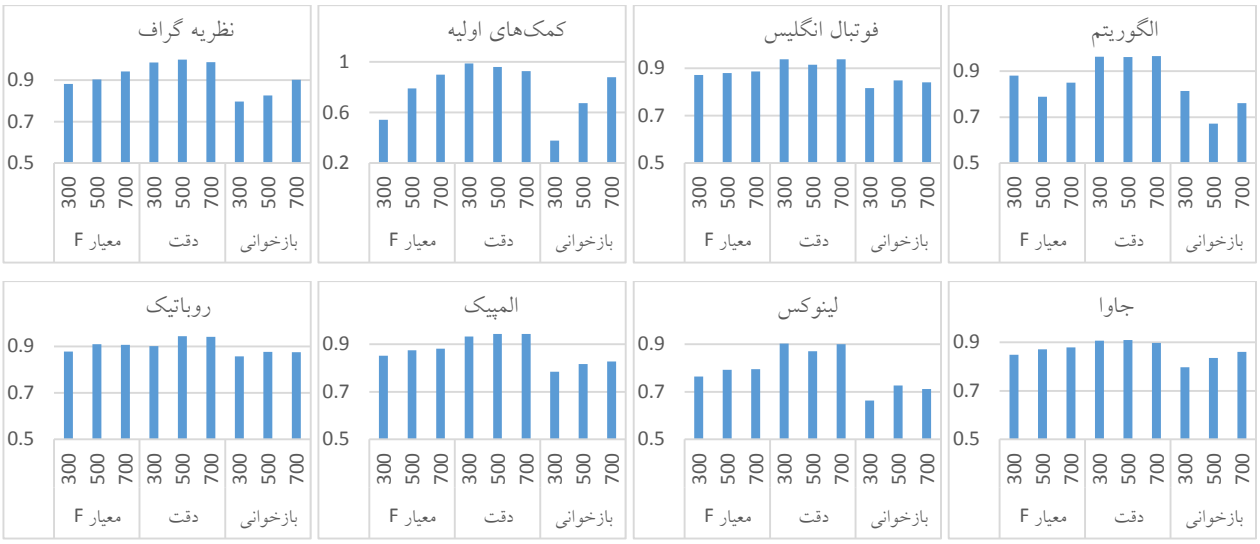
٥-٤. نتایج عملی و تحلیل آن‌ها

٥-٤-١. نتایج مؤلفه پردازش متن

با توجه به تأثیر قابل توجه تعداد ویژگی‌های انتخابی در کارایی رده‌بندهای متن، برای هر موضوع، مجموعه آزمایش‌هایی جهت تعیین تعداد ویژگی‌ها، انجام گرفته است. شکل ۷ نشان دهنده میزان کارایی رده‌بندهای آموزش دیده بر اساس معیارهای بازخوانی، دقت و معیار **F** است. این معیارها بر اساس اعتبارسنجی متقاطع ۱۰ دسته‌ای[1]، محاسبه شده‌اند.

همان‌طور که نمودارها نشان می‌دهند، یک ویژگی بین همه‌ی این رده‌بندها مشترک است و آن دقت بالای آن‌هاست؛ به عبارت

[1] 10-fold cross validation



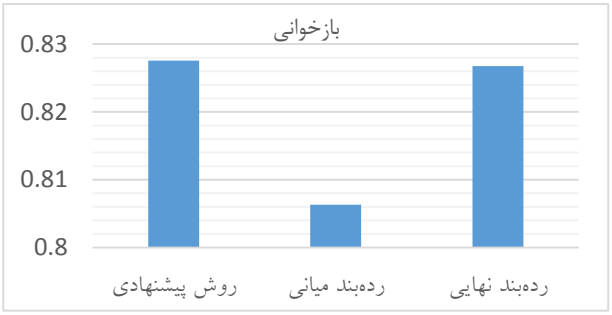
شکل ۷. میزان کارایی رده‌بندهای متن بر اساس معیارهای بازخوانی، دقت و معیار **F**

دیگر اگر نمونه‌ای توسط یکی از این رده‌بندها مربوط تشخیص داده شود، به‌احتمال زیاد این تشخیص صحیح است. در این شرایط، عنصر تعیین کننده‌ی میزان کارایی رده‌بندها بر اساس معیار **F**، دقت آن‌ها است. برای موضوع الگوریتم از ۳۰۰ کلمه و برای دو موضوع کمک‌های اولیه و نظریه گراف از ۷۰۰ کلمه به‌عنوان ویژگی‌های ورودی رده‌بند استفاده شده است. برای سایر موضوعات، استفاده از ۵۰۰ کلمه، به کارایی مطلوب منجر می‌شود.

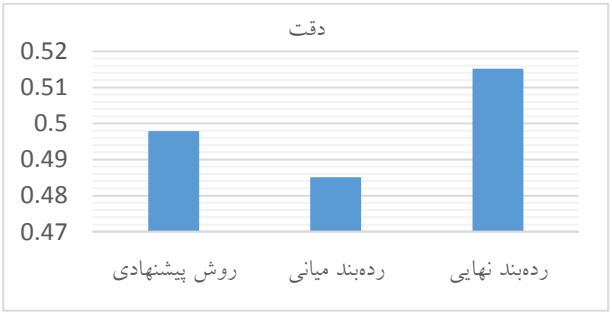
۵-٤-۲. نتایج مؤلفه پردازش تصویر

در این بخش، کارایی روش پیشنهادی را بر اساس معیارهای استاندارد با روش‌های متداول **CNN** عمیق مقایسه می‌کنیم. شکل‌های ۸ و ۹ به ترتیب بر اساس معیارهای بازخوانی و دقت، میزان کارایی روش پیشنهادی را با خروجی **CNN** عمیق بر مبنای رده‌بند نهایی و خروجی **CNN** عمیق بر مبنای رده‌بند میانی، مقایسه می‌کند. در شکل‌ها مشاهده می‌شود که رده‌بند میانی بر اساس هر دو معیار بازخوانی و دقت نسبت رده‌بند نهایی و عملکرد ضعیف‌تری دارد که نشان دهنده تأثیر افزایش عمق و پیچیدگی شبکه در افزایش کارایی شبکه است. روش پیشنهادی بر اساس معیار بازخوانی، از جنبه کارایی تقریباً با رده‌بند نهایی برابر است ولی بر اساس معیار دقت از رده‌بند میانی بهتر و حدود ۱/۵ درصد از رده‌بند نهایی ضعیف‌تر است. با نگاه به شکل ۱۰ در می‌یابیم که این کاهش دقت هم زمان با ثبات بازخوانی، به ازای بهبود ۱۲ درصدی زمان پردازش شبکه به دست آمده است.

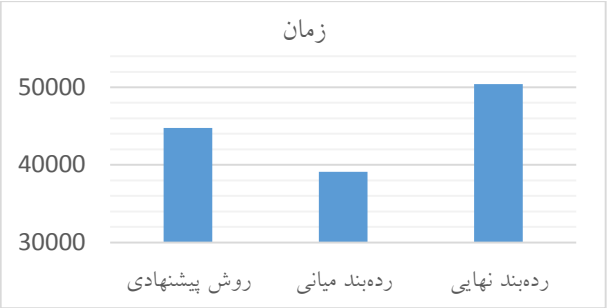
با توجه به این که عمق **CNN** مورد آزمایش ۱۲ است و محاسبات لازم برای خروجی لایه میانی و لایه نهایی تنها ٤ لایه تفاوت دارد، کاهش ۱۲ درصدی زمان اجرا در روش پیشنهادی نسبت به روش نهایی قابل توجه است.



شکل ۸. مقایسه روش پیشنهادی با خروجی رده‌بند میانی و رده‌بند نهایی بر اساس معیار بازخوانی



شکل ۹. مقایسه روش پیشنهادی با خروجی رده‌بند میانی و رده‌بند نهایی بر اساس معیار دقت



شکل ۱۰. مقایسه روش پیشنهادی با خروجی رده‌بند میانی و رده‌بند نهایی بر اساس زمان اجرا برحسب میلی ثانیه

شکل ۱۱. مقایسه روش پیشنهادی با خروجی رده‌بند میانی و رده‌بند نهایی بر اساس میانگین صحت

شکل ۱۱ میزان کارایی روش پیشنهادی را با خروجی‌های میانی و نهایی شبکه بر اساس معیار میانگین صحت، مقایسه می‌کند. همان‌طور که مشاهده می‌شود بر اساس این معیار، کارایی روش پیشنهادی از رده‌بند میانی بالاتر و از رده‌بند نهایی پایین‌تر است. البته بر اساس این معیار، عملکرد روش پیشنهادی بیشتر به رده‌بند نهایی نزدیک است به‌طوری که میانگین صحت روش پیشنهادی حدود ۰/۵ درصد پایین‌تر از رده‌بند نهایی و حدود ۱/۵ درصد بالاتر از رده‌بند میانی است. مقایسه دو شکل ۱۰ و ۱۱ نشان می‌دهد که روش پیشنهادی در مدیریت منابع پردازشی موفق بوده است؛ زیرا همان‌طور که نمودار زمان نشان می‌دهد از نظر زمان تقریباً در جایگاه وسط رده‌بند میانی و نهایی قرار گرفته است اما کارایی آن بر اساس معیار صحت به رده‌بند نهایی بسیار نزدیک‌تر است. در واقع روش پیشنهادی با جلوگیری از ارجاع نمونه‌های ساده به لایه‌های بالاتر شبکه و تعیین برچسب آن‌ها با استفاده از رده‌بند میانی و در نظر گرفتن میزان اطمینان رده‌بند میانی در رابطه با نمونه‌های ورودی، از رده‌بند نهایی تنها برای نمونه‌های دشوارتر استفاده می‌کند و به این ترتیب هزینه کلی فرآیند رده‌بندی تصاویر را کاهش می‌دهد.

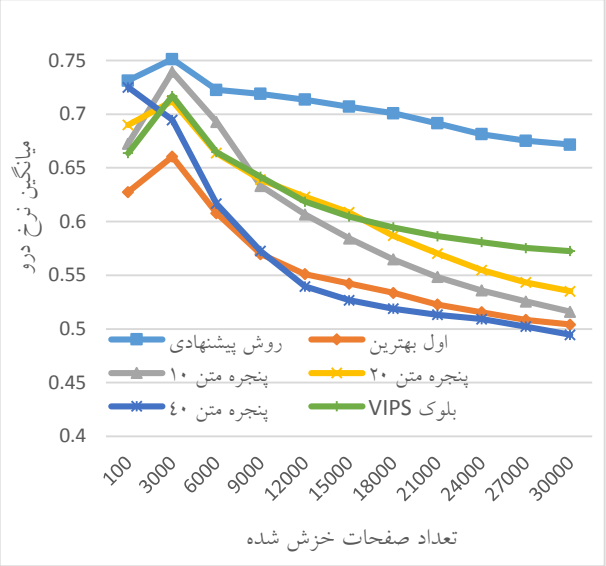
موضوعی که لازم است در به‌کارگیری روش پیشنهادی مدنظر قرار گیرد، برقراری تعادل بین کاهش صحت و افزایش سرعت محاسبات است. هرچه در گره‌های دروازه‌ی روش پیشنهادی، مقدار بالاتری برای حداقل اطمینان جهت توقف فرآیند رده‌بندی در خروجی‌های میانی تعیین شود، رده‌بندی در لایه‌های بالاتر، برای تعداد نمونه‌های بیشتری انجام می‌شود؛ در نتیجه صحت کلی فرآیند رده‌بندی به صحت رده‌بند نهایی مدل پایه نزدیک‌تر می‌شود و سرعت آن در مقایسه با مدل پایه افزایش چندانی ندارد. از طرف دیگر، با کاهش مقدار حداقل اطمینان در گره‌های دروازه، صحت روش پیشنهادی در مقایسه با رده‌بند نهایی مدل پایه کاهش بیشتری دارد، اما در عوض سرعت آن در مقایسه با سرعت رده‌بند نهایی افزایش قابل توجهی می‌یابد. مقدار مناسب حداقل اطمینان در گره‌های دروازه، طی بده‌بستان[1] بین صحت و سرعت مطلوب برای فرآیند، بر اساس مسئله‌ای که مدل در آن به‌کار گرفته شده است، تعیین می‌شود. در آزمایش انجام شده برای پردازش تصاویر، کاهش ۰/۵ درصدی صحت باعث افزایش ۱۲ درصدی سرعت محاسبات می‌شود، که در مسئله‌ی پایش محتوای وب با توجه به حجم زیاد تصاویری که باید پردازش شوند و عدم تغییر قابل توجه در کیفیت فرآیند کلی پایش محتوای وب با کاهش ۰/۵ درصدی صحت، بده‌بستان مذکور بین صحت و سرعت، برای این مسئله مطلوب است. اما ممکن است در مسئله‌ای دیگر، شرایط به نحوی باشد که حفظ صحت فرآیند اهمیت زیادی داشته باشد و افزایش سرعت محاسبات از اولویت پایین‌تری برخوردار باشد؛ به بیان دیگر و از دیدگاه حساس به هزینه، کاهش صحت و افزایش خطا، منجر به افزایش هزینه‌ای در فرآیند کلی شود که بهبود سرعت و کاهش هزینه‌ی محاسبات نتواند آن را جبران کند. در این صورت طی بده‌بستان بین صحت و سرعت فرآیند برای مسئله‌ی مذکور، صحت در اولویت است، که با تعیین مقدار حداقل اطمینان بالاتر در گره‌های دروازه‌ی روش پیشنهادی در زمان اعتبار سنجی مدل، قابل دست‌یابی است. در مسائلی که افزایش سرعت فرآیند

[1] Tradeoff

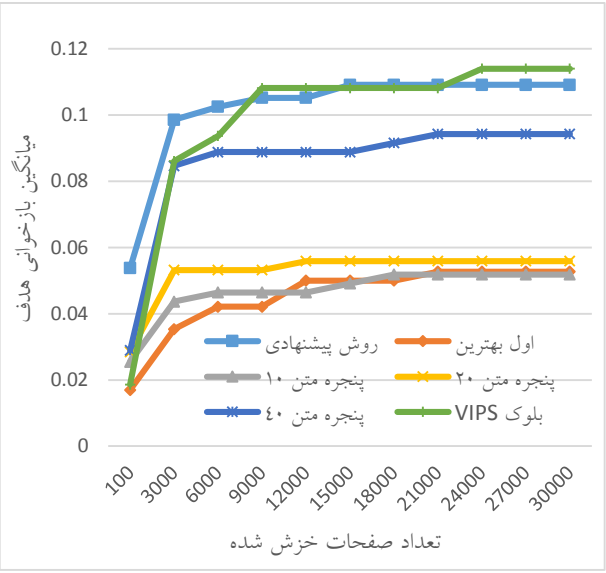
و ایجاد پاسخ با کمترین تأخیر، اهمیت بیشتری نسبت به حفظ صحت فرآیند دارد، می‌توان با کاهش مقدار حداقل اطمینان در گره‌های دروازه، به تعادل مطلوب بین صحت و سرعت دست یافت.

۵-٤-۳. نتایج مؤلفه خزشگر

روش‌های مختلف خزش برای دوره‌ای شامل پایش ۳۰ هزار صفحه‌ی وب بر اساس معیارهای نرخ درو و بازخوانی هدف با یکدیگر مقایسه شده‌اند.



شکل ۱۲. میانگین نرخ درو روش‌ها برای موضوعات مختلف



شکل ۱۳. میانگین بازخوانی هدف روش‌ها برای موضوعات مختلف

همان‌طور که در شکل ۱۲ مشاهده می‌شود نرخ درو روش پیشنهادی با اختلاف قابل توجهی بیشتر از سایر روش‌های مقایسه شده است و بالاترین جایگاه را به خود اختصاص داده است. آن‌چنان که انتظار می‌رود روش بلوک **VIPS** در جایگاه دوم قرار دارد که این موضوع با نتایج گزارش شده در [۱۸] و معرفی بلوک **VIPS** به‌عنوان برترین روش خزش موضوعی در گزارش مذکور منطبق است. در شکل ۱۳ مشاهده می‌شود که بازخوانی هدف روش پیشنهادی و بلوک **VIPS** بسیار به هم نزدیک و در رقابت با یکدیگر هستند و پس از پایش ۳۰ هزار صفحه، اختلاف بسیار اندکی دارند. بر مبنای معیار بازخوانی هدف دو روش مذکور در جایگاه بالاتری نسبت به سایر روش‌ها قرار دارند. البته طبق نتایج ارائه شده در جدول ۵، میزان انحراف معیار نرخ درو و بازخوانی هدف روش پیشنهادی برای موضوعات بررسی شده از روش بلوک **VIPS** و همچنین سایر روش‌های مقایسه شده کمتر است که ثبات بیشتر نتایج گزاش شده برای روش پیشنهادی نسبت به دیگر روش‌ها را نشان می‌دهد.

همچنین نتایج جدول ۵ نشان می‌دهد که روش پیشنهادی بر اساس معیار نرخ درو بعد از دریافت ۳۰ هزار صفحه با اختلاف قابل توجه ۱۰ درصد، عملکرد بهتری نسبت به روش بلوک **VIPS** داشته است و طی مقایسه با نتایج [۱۸] می‌توان گفت روش پیشنهادی به بالاترین سطح نتایج برای خزش موضوعی حساس به هزینه دست یافته است.

جدول ۵. میزان کارایی روش‌ها بعد از خزش ۳۰ هزار صفحه

روش خزش	نرخ درو		بازخوانی هدف	
	میانگین	انحراف معیار	میانگین	انحراف معیار
روش پیشنهادی	۰/٦۷	۰/۱٤	۰/۱۱	۰/۱۰
اول بهترین	۰/۵۰	۰/۳۲	۰/۰۵	۰/۰۵
پنجره متن ۱۰	۰/۵۲	۰/۲۸	۰/۰۵	۰/۰٦
پنجره متن ۲۰	۰/۵٤	۰/۲۷	۰/۰٦	۰/۰۵
پنجره متن ٤۰	۰/٤۹	۰/۲۸	۰/۰۹	۰/۱۰
بلوک **VIPS**	۰/۵۷	۰/۲۸	۰/۱۱	۰/۱۲

٦. جمع‌بندی

بحث هزینه، یک چالش حقیقی در فرماندهی و کنترل مسائل دنیای واقعی محسوب می‌شود، با این حال در پژوهش‌های پیشین به‌طور شایسته مورد مطالعه قرار نگرفته بود. در این مقاله یک بیان رسمی برای فرماندهی و کنترل حساس به هزینه ارائه گردید که بر مبنای آن هزینه کلی فرآیند، بر اساس کارکردهای کلیدی **C4ISR** محاسبه می‌شود. با انتخاب مسئله‌ی پایش محتوای وب به‌عنوان یک نمونه‌ی بااهمیت در حوزه **C4ISR** فضای سایبری، بیان رسمی ارائه شده برای فرماندهی و کنترل حساس به هزینه، به شکل عملی پیاده‌سازی شد. همچنین روش‌هایی نوین برای مؤلفه‌های پردازش تصویر و خزش موضوعی حساس به هزینه ارائه گردید.

روش پیشنهادی پردازش تصویر، که یک روش وفقی حساس به هزینه است، با پردازش سریع نمونه‌های ساده و اختصاص منابع محاسباتی بیشتر به نمونه‌های دشوار، کارایی سامانه را در سطح مطلوب نگه می‌دارد. نتایج آزمایش‌ها نشان می‌دهند روش پیشنهادی پردازش تصویر در مقایسه با روش پایه، در ازای کاهش ۰/۵ درصد از صحت، سرعت محاسبات را ۱۲ درصد افزایش می‌دهد. این روش پیشنهادی به جهت استفاده از رویکرد وفقی در پردازش تصویر حساس به هزینه، از قابلیت ترکیب با سایر روش‌های ایستای حساس به هزینه برای رده‌بندهای **CNN** برخوردار است. در روش پیشنهادی خزش حساس به هزینه انتظار می‌رفت به دلیل استفاده از امتیازهای مجموعه‌ای از روش‌های استخراج زمینه پیوند، به‌جای استفاده تنها از یک روش، پهنای باند به شکل هدفمندتری مورد بهره‌برداری قرار گیرد. نتایج پیاده‌سازی و مقایسه روش خزش پیشنهادی با بالاترین سطح نتایج موجود، نشان دهنده‌ی برتری ۱۰ درصدی روش پیشنهادی بر اساس معیار نرخ درو است.

در آخر، به‌منظور ادامه و تکمیل زنجیره‌های مرتبط با مقاله حاضر دو پیشنهاد ارائه می‌گردد. پیشنهاد اول، پیاده‌سازی **CNN**های عمیق‌تر برای آشکار سازی هرچه بهتر توانمندی‌های روش ارائه شده پردازش تصویر حساس به هزینه است. پیشنهاد دوم، استفاده از بیان ارائه شده برای فرماندهی و کنترل حساس به هزینه برای تمرکز بر بحث هزینه در سایر فرآیندها و سامانه‌های **C4ISR**، مطابق روندی مشابه با آنچه برای **C4ISR** پایش محتوای وب در این مقاله ارائه گردید، است.

منابع و مراجع

[1] M. M. Hurley, "For and from cyberspace: Conceptualizing cyber intelligence, surveillance, and reconnaissance," *Air Sp. Power J.*, vol. 26, no. 6, pp. 12–33, 2012.

[2] J. S. Nye Jr, "Cyber Power." Belfer Center for Science and International Affairs, Harvard Kennedy School, 2010.

[3] J. S. Nye Jr, "Deterrence and dissuasion in cyberspace," *Int. Secur.*, vol. 41, no. 3, pp. 44–71, 2017.

[۴] "سند راهبردی پدافند سایبری کشور،سازمان پدافند غیرعامل کشور," ۱۳۹۴.

[5] D. A. Alberts, D.S., Garstka, J.J., Hayes, R.E., Signori, *Understanding information age warfare*. Assistant secretary of defense (C3I/Command Control Research Program) Washington DC, 2001.

[6] D. S. Alberts and R. E. Hayes, *Understanding Command and Control*. Assistant secretary of defense (C3I/Command Control Research Program) Washington DC, 2006.

[7] M. S. Vassiliou, D. S. Alberts, and J. R. Agre, *C2 Re-envisioned: The Future of the Enterprise*. CRC Press, 2014.

[8] P. Turney, "Types of cost in inductive concept learning," in *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning*, 2000, pp. 15–21.

[9] Z. Jiao, P. Yao, J. Zhang, L. Wan, and X. Wang, "Capability Construction of C4ISR Based on AI Planning," *IEEE Access*, vol. 7, pp. 31997–32008, 2019.

[10] J. Parra-Arnau and C. Castelluccia, "On the cost-effectiveness of mass surveillance," *IEEE Access*, vol. 6, pp. 46538–46557, 2018.

[11] V. Shukla, B. Singh, M. Kumar, and K. Negi, "Big Data Analytics in C4I Systems," in *2018 International Conference on Automation and Computational Engineering (ICACE)*, Oct. 2018, pp. 102–106, doi: 10.1109/ICACE.2018.8687057.

[12] H. Schulze and K. Mochalski, "Internet Study 2008/2009," Ipoque, 2009. [Online]. Available: http://www.ipoque.de/userfiles/file/ipoque-Internet-Study-08-09.pdf.

[13] T. Jo, *Text Mining*, vol. 45. Cham: Springer International Publishing, 2019.

[14] X. Li, Z. Liu, P. Luo, C. Change Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3193–3202.

[15] "https://www.worldwidewebsize.com/," Accessed April 2019.

[16] M. Naghibi and A. T. Rahmani, "Focused crawling using vision-based page segmentation," in *Communications in Computer and Information Science*, 2012, vol. 285, pp. 1–12, doi: 10.1007/978-3-642-29166-1_1.

[17] G. Pant and P. Srinivasan, "Link contexts in classifier-guided topical crawlers," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 107–122, 2006, doi: 10.1109/TKDE.2006.12.

[18] Y. Bin Yu, S. L. Huang, N. Tashi, H. Zhang, F. Lei, and L. Y. Wu, "A survey about algorithms utilized by focused

[33] J.-G. Lee, D. Bae, S. Kim, J. Kim, and M. Y. Yi, "An effective approach to enhancing a focused crawler using Google," *J. Supercomput.*, pp. 1–18, Feb. 2019, doi: 10.1007/s11227-019-02787-9.

[34] C. Iliou, T. Tsikrika, G. Kalpakis, S. Vrochidis, and I. Kompatsiaris, "Adaptive Focused Crawling Using Online Learning," in *Internet Science: 5th International Conference, INSCI 2018*, 2018, pp. 40–53, doi: 10.1007/978-3-030-01437-7_4.

[35] M. Han, P.-H. Wuillemin, and P. Senellart, "Focused Crawling Through Reinforcement Learning," in *ICWE 2018: Web Engineering*, 2018, pp. 261–278, doi: 10.1007/978-3-319-91662-0_20.

[36] M. M. G. Farag, S. Lee, and E. A. Fox, "Focused crawler for events," *Int. J. Digit. Libr.*, vol. 19, no. 1, pp. 3–19, Mar. 2018, doi: 10.1007/s00799-016-0207-1.

[37] P. M. E. De Bra and R. D. J. Post, "Information retrieval in the World Wide Web: Making client-based searching feasible," *Comput. Networks ISDN Syst.*, vol. 27, no. 2, pp. 183–192, 1994.

[38] D. Cai, S. Yu, J. R. Wen, and W. Y. Ma, "VIPS: a visionbased page segmentation algorithm." Microsoft Technical Report, MSR-TR-2003-79, 2003, [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:VIPS:+a+visionbased+page+segmentation+algorithm#0.

[39] A. Kott and D. S. Alberts, "How Do You Command an Army of Intelligent Things?," *Computer (Long. Beach. Calif).*, vol. 50, no. 12, pp. 96–100, 2017, doi: 10.1109/MC.2017.4451205.

[40] G. M. Weiss and Y. Tian, "Maximizing classifier utility when there are data acquisition and modeling costs," *Data Min. Knowl. Discov.*, vol. 17, no. 2, pp. 253–282, 2008.

[41] "http://dmoz-odp.org," Accessed April 2019.

[42] P. Srinivasan, F. Menczer, and G. Pant, "A general evaluation framework for topical crawlers," *Inf. Retr. Boston.*, vol. 8, no. 3, pp. 417–447, 2005.

[43] R. Baeza-Yates, B. Ribeiro-Neto, and others, *Modern information retrieval*, vol. 463. ACM Press, 1999.

[44] "http://htmlparser.sourceforge.net/," Accessed May 2019.

[45] K. S. Jones and P. Willett, *Readings in information retrieval*. Morgan Kaufmann, 1997.

[46] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, pp. 412–420.

[47] "http://www.cs.waikato.ac.nz/ml/weka/index.html," Accessed May 2019.

[48] "http://carl.cs.indiana.edu/fil/IS/JavaCrawlers/," Accessed May 2019.

[49]"http://www.cad.zju.edu.cn/home/dengcai/VIPS/VIPS.html," Accessed August 2019.

[50] "https://caffe.berkeleyvision.org/," Accessed August 2019.

web crawler," *J. Electron. Sci. Technol.*, vol. 16, no. 2, pp. 129–138, 2018, doi: 10.11989/JEST.1674-862X.70116018.

[19] N. S. Board, N. R. Council, and others, *C4ISR for Future Naval Strike Groups*. National Academies Press, 2006.

[20] B. Wilson *et al.*, *Maritime Tactical Command and Control Analysis of Alternatives*. RAND Corporation, 2016.

[21] D. Kehl, K. Bankston, R. Greene, and R. Morgus, "Surveillance Costs: The NSA's Impact on the Economy, Internet Freedom & Cybersecurity," *New Am. Open Technol. Inst.*, no. July, pp. 1–64, 2014, [Online]. Available: http://oti.newamerica.net/sites/newamerica.net/files/policydocs/Surveilance_Costs_Final.pdf.

[22] A. Polyak and L. Wolf, "Channel-Level Acceleration of Deep Face Representations," *Access, IEEE*, vol. 3, pp. 2163–2175, 2015.

[23] J. Ba and R. Caruana, "Do deep nets really need to be deep?," in *Advances in neural information processing systems*, 2014, pp. 2654–2662.

[24] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015, pp. 1–13.

[25] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating Very Deep Convolutional Networks for Classification and Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 1943–1955, Oct. 2016.

[26] N. Vasilache, J. Johnson, M. Mathieu, S. Chintala, S. Piantino, and Y. LeCun, "Fast convolutional nets with fbfft: A GPU performance evaluation," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015, pp. 1–17.

[27] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on CPUs," in *Proc. Deep Learning and Unsupervised Feature Learning NIPS Workshop*, 2011, vol. 1.

[28] M. Courbariaux, Y. Bengio, and J.-P. David, "Low precision arithmetic for deep learning," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015, pp. 1–10, doi: arXiv: 1412.7024.

[29] V. N. Murthy, V. Singh, T. Chen, R. Manmatha, and D. Comaniciu, "Deep decision network for multi-class image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2240–2248.

[30] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325–5334.

[31] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[32] M. Klein, L. Balakireva, and H. Van de Sompel, "Focused Crawl of Web Archives to Build Event Collections," in *Proceedings of the 10th ACM Conference on Web Science - WebSci '18*, 2018, pp. 333–342, doi: 10.1145/3201064.3201085.