مجله علمی- پژوهشی "پدافند الکترونیکی و سایبری"
سال پنجم، شماره ۴، زمستان ۱۳۹۶، ص ۹۳-۸۱

تشخیص ربات‌های ناهنجار در پرس‌وجوهای موتور جستجو

محمدجواد سروقدمقدم[۱]، مهدی نقوی[۲*]، مجید غیوری ثالث[۳]

۱- دانشجوی کارشناسی ارشد، ۲- استادیار، ۳- استادیار، گروه کامپیوتر دانشگاه جامع امام حسین (ع)

(دریافت:۹۵/۰۵/۲۶، پذیرش: ۹۶/۰۵/۰۱)

چکیده

موتورهای جستجو را می‌توان بهترین ابزار کارآمد برای مدیریت، بازیابی و استخراج اطلاعات مهم از مجموعه عظیم داده‌های وب معرفی کرد. این موتورها پهنه وسیع وب را به‌طور زمان‌بندی‌شده پیمایش می‌کنند و به جمع‌آوری صفحات بی‌شمار ذخیره‌شده در گوشه کنار وب می‌پردازند. ارائه‌دهندگان موتورهای جستجو همواره به دنبال بهبود ارتباط نتایج و کاهش زمان پاسخ به کاربران هستند، اما هر دو این موارد می‌تواند تحت تأثیر ترافیک خودکار ارسال‌شده از سوی ربات‌ها قرار گیرد. در این مقاله ابتدا به تعریف ربات‌ها و چالش تشخیص آن‌ها پرداخته شده است. سپس، روشی با نام بوف برای تشخیص ربات‌های جستجو ارائه شده است. در روش بوف برای دستیابی به دقتی بالا در تشخیص ربات‌های ناهنجار، از پارامترهای مختلف و نسبتاً زیادی برای مدل‌کردن رفتار کاربران استفاده شده است. پس از تعیین اولویت پارامترها در تشخیص ماهیت کاربران، درخت تصمیمی ساخته شده و اقدام به دسته‌بندی کاربران در گروه‌های انسان، ربات مخرب، ربات مجاز و نامشخص می‌کند. ربات‌های تشخیص داده‌شده در درخت تصمیم، بخش دیگری از سامانه تشخیص ربات را فعال می‌کند که قادر است با توجه به الگوی رفتاری شبکه‌های رباتی، حتی ربات‌هایی با نرخ درخواست پایین را نیز شناسایی کند. ارزیابی روش پیشنهادی بر روی داده‌های آزمون، صحت ۹۷/۷ درصدی را در تشخیص ماهیت کاربران نشان می‌دهد که حداقل بهبود دقت ۹/۹ درصدی را نسبت به روش‌های بررسی‌شده در این حوزه نشان می‌دهد. رقم قابل توجهی که در هر روز تصمیم‌گیری در مورد ۲۲۳۰ کاربر را تحت تاثیر قرار می‌دهد.

واژه‌های کلیدی: موتور جستجو، ربات جستجو، تحلیل داده ثبت رویداد، تشخیص ربات، درخت تصمیم

۱- مقدمه

امروزه می‌توان از وب جهان‌گستر، به‌عنوان بهترین محیط برای تولید، انتشار و دسترسی به دانش نام برد که به دلیل حجم زیاد اطلاعات، ناهمگنی، رشد نمایی و عدم ساختار مناسب، به‌طور مرتب نیاز به ابزارها، روش‌ها و راه‌بردهای جدید خودنمایی می‌کند. در این میان موتورهای جستجو را می‌توان بهترین ابزار برای دسترسی به این محیط معرفی کرد. با توجه به این‌که حدود ۸۰ ٪ از کاربران از طریق موتورهای جستجو به سایت‌ها و اطلاعات مورد نظرشان دسترسی پیدا می‌کنند [۱]، بررسی پتانسیل موتورهای جستجو از اهمیت بالایی برخوردار است. با گذشت زمان نیز تعداد کاربرانی که از موتورهای جستجو استفاده می‌کنند به‌سرعت درحال افزایش است. مطابق آمار، موتورهای جستجوی پرکاربرد دنیا میلیاردها صفحه را تحت پوشش قرار داده و چندین میلیارد پرس‌وجو در روز دریافت می‌کنند [۱]. طبیعی است که برای پاسخگویی این حجم عظیم از پرس‌وجوها با چالش‌های جدیدی مواجه خواهیم بود. حجم داده‌های ذخیره‌شده به‌سرعت درحال رشد است و همواره نیاز به افزایش حجم ذخیره‌سازها، در عین حفظ کارایی وجود دارد؛ در سمت دیگر اندازه، سرعت قرار دارد به‌نحوی‌که هر چه حجم مجموعه داده برای پردازش بیشتر باشد، زمان بیشتری برای پردازش صرف می‌شود و از طرفی موتورهای جستجو نیاز به ارائه نتایج این پردازش‌ها به‌صورت بی‌درنگ دارند [۲]. استفاده ناهنجار برخی کاربران از دیگر چالش‌های موتور جستجو می‌باشد. مزایایی که با دست‌کاری در نتایج حاصل از جستجو به دست می‌آید، باعث شده است که سوءاستفاده‌های گسترده‌ای در موتورهای جستجو مشاهده شود که استفاده از ربات‌ها یکی از مهم‌ترین این موارد می‌باشد.

پرس‌وجوی تولیدشده توسط ربات‌ها منجر به پرشدن ظرفیت پردازشی موتور جستجو به‌ویژه در زمان‌های اوج مصرف می‌شود و در نتیجه، موتور جستجو را در پاسخگویی به کاربران واقعی خود با مشکل روبرو می‌کند (افزایش زمان پاسخ و یا حتی ناتوانی در پاسخگویی) که این مسئله نارضایتی کاربران را به همراه دارد. علاوه بر این، داده‌های ثبت رویداد در موتور جستجو، تصویری از چگونگی تعامل کاربران با دنیای اینترنت است و اغلب به‌منظور انجام تحلیل‌های متنوع مورد استفاده قرار می‌گیرند. واضح است که فعالیت‌های بدخواهانه ربات‌ها به‌منظور شیوع برخی پرس‌وجوهای

* رایانامه نویسنده مسئول: mnaghavi@ihu.ac.ir

خاص منجر به ایجاد انحراف در نتایج این‌گونه تحلیل‌ها خواهد شد. مورد سومی که می‌تواند برای موتور جستجو مضر باشد این است که الگوریتم‌های رتبه‌بندی استفاده‌شده در برخی از موتورهای جستجو متأثر از نحوه تعامل کاربران با سامانه است به این صورت که نتایجی که بیشتر انتخاب می‌شوند ممکن است در رتبه‌بندی ارزش بالاتری بگیرند، در نتیجه این ربات‌ها می‌توانند رتبه نتایج دلخواه خود را به‌صورت مصنوعی افزایش دهند [۳]. از این جهت، تعیین پارامترهایی برای تمایز بین ترافیک تولیدشده توسط ربات‌ها و فعالیت‌های جستجو کاربران واقعی، امری ضروری به نظر می‌رسد. با این حال، تعیین چنین پارامترهایی و در نتیجه تشخیص ربات‌ها با مشکلات زیادی همراه است.

مهاجمان برای پوشش فعالیت‌های خود از روش‌های پیچیده‌ای مانند تقلید رفتارهای انسانی و حتی ربودن ترافیک‌های قانونی کاربران استفاده می‌کنند. علاوه بر این، بعد از انجام اقدامات متقابل، با تغییر شیوه‌های رفتاری خود شروع به تکامل و مقابله با فعالیت‌های انجام‌شده می‌کنند [۴].

۲- معرفی ربات‌های جستجو

ربات‌ها برنامه‌هایی هستند که با ارسال پرس‌وجوها و کلیک‌کردن خودکار، می‌توانند بخش عمده‌ای از ظرفیت پردازشی موتور جستجو را مصرف کنند. این ربات‌ها با اهداف گوناگونی در حال تعامل با موتور جستجو هستند که بخشی از آن‌ها را ربات‌های متخاصم تشکیل می‌دهند. در نگاه نخست ربات‌های متخاصم با ارسال پرس‌وجوهای زیاد از یک طرف، موجب ایجاد ازدحام زیاد در موتور جستجو و افزایش میانگین زمان پاسخ خواهند شد و از طرف دیگر، شیوع برخی پرس‌وجوهای خاص و افزایش مصنوعی رتبه برخی نتایج را به همراه دارند [۵].

گروهی از این ربات‌ها اقدام به ارسال تعداد زیادی پرس‌وجو از روی یک فرهنگ لغات می‌کنند و در نتیجه، کلمات ارسال‌شده تنها در یک یا دو حرف با یکدیگر فاصله دارند [۶]. گروهی دیگر نیز وجود دارند که به‌طور مکرر با جستجوی کلمات کلیدی در یک زمینه خاص و انتخاب نتایج مورد نظر خود، سعی در بهبود رتبه آن صفحات در فهرست نتایج موتور جستجو دارند [۷]. گونه دیگری از ربات‌ها وجود دارند که سعی دارند با انجام مهندسی معکوس محتوای سایت‌های برتر در حوزه خود را به‌دست آورده و در وب‌سایت‌های خود استفاده کنند [۸]. در مواجه با صفحاتی که این ربات‌ها را در اختیار گرفته‌اند، این مسئله به چشم می‌خورد که تنها فهرستی از کلمات کلیدی پرکاربرد محتوای آن‌ها را تشکیل می‌دهد و تمرکز اصلی آن‌ها به نمایش گسترده تبلیغات مختلف است. علاوه‌بر این، موارد ربات‌هایی نیز وجود دارند که با هدف افزایش رتبه برخی کلمات کلیدی عمدتاً غیراخلاقی، در تلاش برای ایجاد اختلال در سامانه پیشنهاددهنده موتور جستجو هستند. این ربات‌ها کلمات کلیدی موردنظر خود را هزاران بار در روز به‌سوی سرویس‌دهنده‌های موتور جستجو ارسال می‌کنند و در مقابل هیچ یک از نتایج بازگردانده‌شده را انتخاب نخواهند کرد [۹]. اما ربات‌ها همیشه به‌صورت یک ماشین تنها فعالیت خود را انجام نمی‌دهند و در برخی موارد ربات‌ها به‌منظور ایجاد حملاتی گسترده و همچنین جلوگیری از شناسایی‌شدن شبکه‌ای را تشکیل داده و در قالب آن اقدام به انجام فعالیت‌های بدخواهانه خود می‌کنند. یک شبکه رباتی[1] شامل گروهی از ماشین‌ها می‌شود که با همکاری یکدیگر دستوراتی را از سرویس‌دهنده فرماندهی و کنترل[2] دریافت کرده و به آن‌ها پاسخ می‌دهند. سرویس‌دهنده فرماندهی و کنترل خود به‌عنوان سازوکاری برای خدمت‌دهی به یک ربات ارشد[3] که معمولاً تحت‌نظر یک کنترل‌کننده انسانی است به‌کار گرفته می‌شود [۱۰]. البته همیشه این ماشین‌ها از اقداماتی که انجام می‌دهند مطلع نیستند و در بیشتر موارد آن‌ها نیز قربانی این شبکه‌ها هستند.

البته در این میان ربات‌های غیرمخربی نیز وجود دارند. این دسته از ربات‌ها به‌صورت کاملاً خوش‌رفتار برخورد کرده و معمولاً خود را از همان ابتدا معرفی می‌کنند، مانند خزشگرهای سایر موتورهای جستجو و یا سامانه‌هایی که به‌منظور تعیین رتبه وب‌سایت‌ها طراحی و استفاده می‌شوند. این ربات در قسمت عامل کاربر از بسته ارسالی، خود را معرفی می‌کنند. ربات‌های متعلق به فراموتورهای جستجو نیز مثال دیگری از این‌گونه ربات‌ها هستند. این ربات‌ها پرس‌وجوهای متعلق به کاربران واقعی را دریافت کرده، به‌سوی موتورهای جستجو ارسال می‌کنند و نتایج به‌دست‌آمده را برای کاربران خود نمایش می‌دهند.

۳- پیشینه تحقیق

در چند سال گذشته به‌علت گسترش روزافزون صفحات وب، موتورهای جستجو به ابزاری ضروری برای استفاده از این صفحات تبدیل شده و نقش مهمی را در زندگی روزانه ما ایفا می‌کنند. اما متأسفانه بخش مهمی از پرس‌وجوهای ارسال‌شده به موتورهای جستجو به‌وسیله ربات‌ها صورت می‌گیرد. تحقیقات انجام‌شده توسط Zhang و همکاران [۸] نشان می‌دهد که حداقل ۴/۱۶ ٪ از کاربران موتورهای جستجو را ربات‌ها تشکیل می‌دهند و همچنین قابل ذکر است که حدود ۳۳ ٪ از ترافیک خودکار ارسال‌شده توسط آن‌ها با

1- Botnet
2- Command and control server (C&C server)
3- Botmaster

هدف آسیب‌زدن به سامانه و ۱۱٪ نیز به‌منظور دستیابی به اطلاعات حساب سایر کاربران ارسال شده‌اند. این آمار باعث شده است که توجه روزافزونی به تشخیص ربات‌ها در پرس‌وجوهای موتورهای جستجو معطوف گردد. نخستین کارهای انجام‌گرفته در حوزه تشخیص ربات‌های جستجوی فعال در موتورهای جستجو تنها پارامترهای ساده‌ای را مورد بررسی قرار می‌دهند. Buzikashvili [۱۱] کاربرانی که در پنجره‌های زمانی مشخص تعداد زیادی پرس‌وجو را ارسال کرده‌اند به‌عنوان ربات درنظر گرفته است. Zhang و همکاران [۱۲] کاربرانی را که بر روی هیچ یک از نتایج جستجو کلیک نکرده‌اند و Jansen و همکاران [۱۳] کاربران با بیش از ۱۰۰ پرس‌وجوی موفقیت‌آمیز را ربات شناسایی کرده‌اند. Sadagopan و Li [۵] کاربرانی که جریان کلیک آن‌ها یک ترتیب منطقی از وقایع را پیروی می‌کنند به‌عنوان کاربران معمولی[1] شناسایی کرده و درمقابل جریان‌های کلیکی که ترتیبی منطقی از وقایع را نشان نمی‌دهند، متعلق به ربات‌ها می‌دانند. برای این منظور، ابتدا تمام فعالیت‌هایی که کاربران می‌توانند انجام دهند، تعریف شده است. سپس با توجه به فعالیت‌های تعریف‌شده هریک از نشست‌های کاربران با یک زنجیره مارکوف مدل می‌شود. پس از آن با توجه به احتمال تخصیص داده‌شده به هر یک از انتقال‌ها، به ازای هر نشست امتیاز حداکثر احتمال[2] محاسبه می‌شود. پس از این مرحله با استفاده از فاصله ماهالانوبیس[3]، موارد دارای فاصله زیاد از مقدار میانگین به‌عنوان ترافیک نامعمول[4] شناسایی می‌شود.

Kang و همکاران [۶] ابتدا ربات‌های جستجو را معرفی نموده و به بیان برخی پارامترهای عددی برای تشخیص ربات‌ها پرداخته است. سپس الگوریتم تشخیص ربات خود را که الگوریتمی نیمه‌نظارتی می‌باشد، بیان نموده است. نیمه‌نظارتی به این معنا که الگوریتم بیان‌شده برای اجراشدن نیاز به مجموعه داده‌ای دارد که بخشی از داده‌های آن (داده‌های ثبت رویداد) باید از ابتدا به‌عنوان ربات و یا انسان علامت‌گذاری شده باشند و سپس این الگوریتم با توجه به ویژگی‌های آن‌ها به تشخیص بخش دیگری از کاربران بپردازد. Daswani و Stoppelman [۱۴] به تشریح یک شبکه بزرگ از ربات‌های جستجو پرداخته‌اند و جزئیات عملکردی و فنی آن‌را به تفصیل بیان کرده‌اند. Stokes و همکاران [۱۵] ابتدا نمونه‌ای از چگونگی رفتار یک ربات و میزان درخواست‌هایی که برای سامانه ارسال می‌کند، آورده‌اند. در بخش بعدی به بیان ویژگی‌های برخی ربات‌های شناخته‌شده در این حوزه پرداخته شده است و اهداف آن‌ها از تعامل با سامانه را بیان می‌کند. این مقاله پارامترهای موردنظر خود را برای تمایز ربات‌ها از انسان‌ها در قالب دو گروه پارامترهای فیزیکی و پارامترهای رفتاری بیان نموده است. در انتها نیز کارایی سامانه تشخیص ربات ارائه‌شده با استفاده از برخی روش‌های نظارتی بر روی داده‌های برچسب‌گذاری‌شده با استفاده از نرم‌افزار داده‌کاوی وکا ارزیابی شده است. Kitts و همکاران [۷] ابتدا به بررسی و بیان ویژگی‌های چند نمونه از ربات‌های جستجو پرداخته‌اند و سپس معیارهایی عددی برای تشخیص و شناخت ربات‌ها عنوان شده است. در ادامه برای تشخیص ربات‌های جستجو و ربات‌های تبلیغاتی، از روش کلاس‌بندی درخت تصمیم استفاده شده است. سپس در جدولی تأثیر هرکدام از این معیارها را در تعیین ربات‌های معرفی‌شده نشان داده است. در این مقاله بیان شده است که حتی موتور جستجوی گوگل هم اطلاعات کاملی ندارد که مشخص کند کدام کلیک‌ها معتبر و کدام‌یک نامعتبر هستند و بنابراین، تعیین نرخ کارایی فیلترهای به‌کار گرفته‌شده بدون این اطلاعات، غیرممکن است. Duskin و Feitelson [۱۶] هدف خود از فیلترکردن فعالیت‌های خودکار در موتور جستجو را اولاً به‌منظور ارائه یک تصویر قابل اطمینان‌تر از فعالیت‌های کاربران انسانی و سپس فراهم‌کردن مبنایی برای ارزیابی تأثیر حضور فعالیت‌های خودکار در موتور جستجو مطرح کرده‌اند. همچنین نویسنده نبود کارهایی قابل‌قبول در این حوزه را شگفت‌آور می‌داند. در ادامه روش پیشنهادی خود که روشی مبتنی بر قانون است را مطرح می‌کند. Yasmin و همکاران [۱۷] به منظور شناخت الگوی رفتاری انسان‌ها و ربات‌ها، ابتدا روشی برای تشخیص ربات‌ها در سرویس‌دهنده‌های وب معرفی کرده اند. در این روش از پنج پارامتر استفاده شده است. پس از شناخت ربات‌ها براساس پارامترهای تعریف‌شده، به بررسی الگوی رفتاری کاربران از نقطه‌نظر تعداد دسترسی‌ها، درخواست صفحات یکسان، دریافت صفحات در بازه‌های زمانی یکسان و فهرست صفحات دریافت‌شده، پرداخته شده است. Srivastava و همکاران [۱۸] در مقاله خود به بررسی روش‌های تشخیص ربات در سرویس‌دهنده‌های وب پرداخته‌اند. برای این منظور، چهار روش متفاوت برای تشخیص ربات‌های قانونی را به‌کار گرفته و با یکدیگر مقایسه کرده‌اند. این روش‌ها شامل بررسی آدرس شبکه ربات‌های شناخته‌شده، بررسی دسترسی به فایل متنی ربات، نگاشت عامل کاربر و جستجوی کلمات کلیدی در عامل کاربر است. بررسی‌های این پژوهشگران نشان می‌دهد که بررسی آدرس شبکه کمترین تاثیر و جستجوی کلمات کلیدی در عامل کاربر بیشترین تاثیر را در تشخیص ربات‌های قانونی دارد. Dong و همکاران [۱۹] برای تشخیص ربات‌های سرویس‌دهنده‌های وب، یازده خصیصه را برای هر کاربر محاسبه کرده‌اند. درگام بعد ابتدا کاربران با تعداد درخواست‌های زیاد را به عنوان ربات برچسب‌گذاری کرده و سپس با استفاده از ماشین بردار پشتیبان نیمه نظارتی اقدام به برچسب‌گذاری سایر کاربران می‌کند. در نهایت، نیز روش پیشنهادی با کمک معیارهای

1- Typical users
2- Maximum likelihood score
3- Mahalanobis distance
4- Atypical

دقت، فراخوانی و F1 با برخی روش‌های مطرح در داده‌کاوی از جمله درخت تصمیم، شبکه عصبی و نزدیکترین همسایه مقایسه شده است. تمام مقالات بررسی شده در این بخش به همراه حوزه فعالیت و روش اتخاذ شده، در جدول (۱) آمده است.

جدول (۱): مهمترین کارهای بررسی‌شده در حوزه تشخیص ربات وب

نویسنده	حوزه فعالیت	رویکرد	روش
Buzikashvili [۱۱]	موتور جستجو	نظارتی	مبتنی بر قانون
Zhang و همکاران [۱۲]	موتور جستجو	نظارتی	مبتنی بر قانون
Jansen و همکاران [۱۳]	موتور جستجو	نظارتی	مبتنی بر قانون
Sadagopan و لی [۵]	موتور جستجو	بدون نظارت	زنجیره مارکوف
Stokes و Buehrer [۱۵]	موتور جستجو	نظارتی	روش‌های کلاس‌بندی
Duskin و Feitelson [۱۶]	موتور جستجو	نظارتی	مبتنی بر قانون
Kang و همکاران [۶]	موتور جستجو	نیمه نظارتی	شبکه بیزین
Kitts و همکاران [۷]	موتور جستجو	نظارتی	درخت تصمیم
Yasmin و همکاران [۱۷]	سرویس‌دهنده‌های وب	نظارتی	مبتنی بر قانون
Srivastava و همکاران [۱۸]	سرویس‌دهنده‌های وب	نظارتی	مبتنی بر قانون
Dong و همکاران [۱۹]	سرویس‌دهنده‌های وب	نیمه نظارتی	ماشین بردار پشتیبان

بررسی مقالات انجام‌گرفته در این حوزه نشان می‌دهد که از میان کارهای انجام‌شده در موتورهای جستجو، تشخیص ربات‌های تبلیغاتی بسیار مورد توجه بوده است. همچنین، در تمام روش‌های پیشنهاد شده، کاربرانی که در یک جلسه رفتاری مشابه ربات مانند نرخ ارسال بالا، حجم زیاد درخواست‌ها، ارسال در بازه‌های زمانی یکسان و غیره را از خود نشان می‌دهند به عنوان ربات درنظر گرفته می‌شوند. این مساله باعث می‌شود که ربات‌ها با شناخت این روش‌های تشخیص، برای دستیابی به اهداف خود از نرخ ارسال و تعداد درخواست‌های کمتر و اما تعداد جلسات بیشتری استفاده کنند. بنابراین، نبود روشی که بتواند همزمان ربات‌های جستجوی به صورت یک ماشین تنها و با رفتار غیرمتعارف و نیز ربات‌های جستجویی که به صورت شبکه‌ای و با نرخ پایین‌تر، سرویس‌دهنده‌های موتور جستجو را تحت تاثیر قرار داده‌اند را تشخیص دهد، به یکی از خلأهای پژوهشی در این حوزه تبدیل شده است.

۴- روش پیشنهادی

روش تشخیص ربات تشریح‌شده در این پژوهش فعالیت خود را با دریافت داده‌های ثبت رویداد در موتور جستجو آغاز می‌کند. براساس محتوای ثبت‌شده در این داده‌ها، صفاتی به منظور تحلیل رفتار هر کاربر استخراج می‌شود. سپس رفتار کاربران براساس صفات استخراج‌شده، مدل می‌شود. به منظور ساخت مدل کلاس‌بندی ابتدا نیاز به داده‌های آموزشی وجود دارد. با تولید داده آموزشی می‌توان اعتبار هریک از صفات توصیف‌شده را مورد سنجش قرار داد و از بین آن‌ها بهترین صفات را برگزید. پس از این مرحله، امکان ساخت مدل کلاس‌بندی موردنظر فراهم خواهد شد. برای این‌که بتوان عملکرد روش ارائه‌شده را ارزیابی نمود، باید مجموعه داده آزمون در اختیار باشد. درصورتی‌که برای آموزش مدل کلاس‌بندی و ارزیابی آن از داده‌های یکسانی استفاده شود و یا حتی شیوه تولید این داده‌ها یکسان باشد، نتایج به‌دست‌آمده خیلی قابل اعتماد نخواهد بود، به همین علت، در این پژوهش تولید داده آموزشی و داده‌های آزمون به روش‌هایی متفاوت و در گام‌های مختلف صورت می‌گیرد. شکل (۱) چارچوب کلی روش پیشنهادی در این پژوهش را نشان می‌دهد.

با توجه به ویژگی‌های منحصر به فردی که بوف (جغد) در تشخیص و شکار صید خود دارد، از این پس روش پیشنهادی خود را با نام روش بوف می‌نامیم.

۴-۱- تحلیل مجموعه داده

برای این‌که قادر به تعیین پارامترهایی در تشخیص ربات‌ها باشیم، نیاز به جمع‌آوری داده‌هایی از بسته‌های ارسال‌شده از طرف کاربران به موتور جستجو و نیز پاسخ‌هایی که از طرف سرورهای موتور جستجو به کاربران ارسال می‌شود، وجود دارد. با وجود این‌که تمام موتورهای جستجو داده‌های ثبت رویداد حاصل از تعامل کاربران را جمع‌آوری می‌کنند اما متأسفانه تعداد بسیار کمی از این داده‌های ثبت رویداد برای فعالیت‌های تحقیقاتی منتشر شده است. مجموعه داده‌ای که در این مقاله مورد تحلیل و بررسی قرار گرفته است، داده‌های ثبت رویداد مربوط به موتور جستجوی بومی یوز می‌باشد. این مجموعه داده حدود یک میلیون داده ثبت رویداد را شامل می‌شود که حاصل از ارتباط ۲۲۵۳۰ کاربر با این موتور جستجو است. از این مجموعه داده ثبت رویداد، هشت فیلد مهم قابل استخراج است که در ادامه به آن‌ها اشاره می‌شود:

۱- آدرس فرستنده: این فیلد شناسه‌ای را برای هر کاربر تعیین می‌کند و برای تمایز کاربران می‌تواند مورد استفاده قرار بگیرد. در حقیقت، هر ماشین در این مجموعه داده با یک آدرس یکتا هویت پیدا می‌کند.

۲- زمان: این فیلد زمان تعامل کاربر با سامانه را نمایش می‌دهد و به کمک آن می‌توان به تعیین پارامترهای زیادی از جمله نرخ ارسال درخواست‌ها پرداخت.

۳- کد وضعیت پاسخ: موفقیت‌آمیزبودن یا بروز خطا در پاسخگویی به درخواست کاربر در این فیلد مشخص می‌شود.

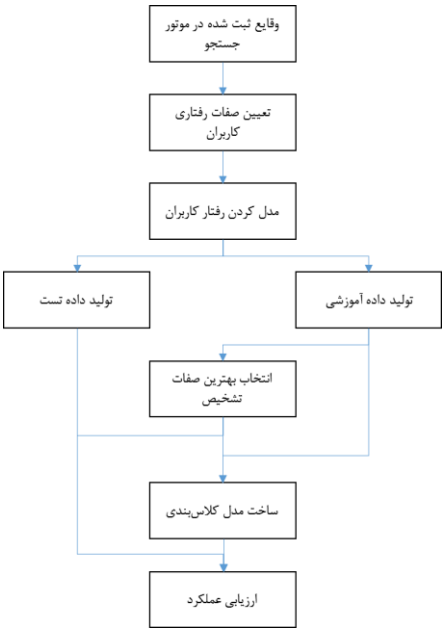
۴- نوع فعالیت: نوع فعالیت کاربر که شامل تایپ‌کردن، جستجو کردن، انتخاب نتایج و کلیک بر روی پیوندهای سایت است را مشخص می‌کند.

۵- پرس‌وجو: متن درخواست کاربر در این قسمت نمایش داده می‌شود.

۶- رتبه نتیجه انتخاب‌شده: این فیلد بیان‌کننده نتایجی است که توسط کاربر به ازای یک پرس‌وجوی خاص انتخاب می‌شود. در این قسمت از داده ثبت رویداد رتبه صفحات انتخاب‌شده ثبت خواهد شد. بدین ترتیب تعداد کلیک‌های کاربر نیز از طریق این داده‌ی ثبت رویداد قابل بازیابی است. از طریق این فیلد می‌توان به نرخ کلیک کاربران و تعداد کلیک‌های آن‌ها به ازای کلمات کلیدی مشخص دست یافت.

۷- آدرس پیوند انتخاب‌شده: درصورتیکه کاربر نتیجه‌ای را از فهرست نتایج نمایش داده شده، انتخاب کرده باشد در این فیلد آدرس صفحه مورد نظر مشخص می‌شود.

۸- عامل کاربر: سیستم‌عامل و مرورگری که کاربر با استفاده از آن‌ها درخواست خود را ارسال کرده است در این قسمت از داده‌ی ثبت رویداد مشخص می‌شود.



شکل (۱): چارچوب کلی روش پیشنهادی (روش بوف)

۴-۲- تعیین صفات رفتاری و مدل کردن رفتار کاربران

به‌منظور مدل‌کردن رفتار کاربران و درنهایت تصمیم‌گیری در مورد ماهیت آن‌ها نیاز به تعیین صفاتی برای هر کاربر وجود دارد. این صفات، پروفایلی را برای هر کاربر تشکیل می‌دهد. در ادامه این صفات معرفی خواهند شد.

تعداد پرس‌وجو: یکی از قوی‌ترین پارامترها در تشخیص ربات‌ها، حجم پرس‌وجوها می‌باشد. ربات‌ها معمولاً تعداد درخواست‌های بسیار بیشتری را در مقایسه با یک فرد معمولی ارسال می‌کنند. درحالی‌که ارسال بیش از ۲۰۰ پرس‌وجو برای کاربران انسانی در یک روز امکان‌پذیر می‌باشد، اما این امر به‌ندرت اتفاق می‌افتد. بررسی‌ها نشان می‌دهد که کاربران با بیش از این تعداد پرس‌وجو را عمدتاً ربات‌ها تشکیل می‌دهند [۲۰].

نرخ ارسال درخواست: این پارامتر حداکثر تعداد درخواست‌های ارسال شده در واحد زمان می‌باشد. ربات‌ها معمولاً با نرخ ارسال بالاتری نسبت به کاربران معمولی پرس‌وجو ارسال می‌کنند. برای تمایز ربات‌ها از کاربران انسانی می‌توان از میانگین و یا حداکثر میزان این پارامتر استفاده نمود. مطالعات انجام‌شده توسط Buehrer و همکاران [۲۰] نشان می‌دهد کاربران انسانی به‌ندرت بیشتر از هفت درخواست را در بازه‌های زمانی ۱۰ ثانیه‌ای ارسال می‌کنند.

مدت زمان فعالیت: فعال‌بودن کاربر به مدت زیاد و به‌طور پیوسته، احتمال ربات‌بودن وی را افزایش می‌دهد. برای مثال، اگر کاربری بیش از ۱۰ ساعت به‌طور پیوسته فعال باشد، به احتمال بیشتری ربات خواهد بود.

نظم در ارسال پرس‌وجو: احتمال ربات‌بودن کاربرانی که یک پرس‌وجوی یکسان را در تعداد دفعات زیاد ارسال می‌کنند و یا کاربرانی که دقیقاً طی یک برنامه زمان‌بندی‌شده ارسال داده دارند زیاد است [۱۶].

نرخ کلیک کاربر: ربات‌ها معمولاً با هدف دستیابی به اطلاعات و یا بررسی شاخص‌های موتور جستجو به‌کار گرفته می‌شوند، بنابراین، تعداد کلیک کمتری نسبت به کاربران معمولی دارند. کارهای گذشته در این حوزه نشان می‌دهد کاربران انسانی در کمترین حالت به ازای هر ده پرس‌وجو، حداقل یک‌بار کلیک خواهند کرد. این مسئله نشان می‌دهد احتمال ربات‌بودن کاربرانی که هرگز کلیک نمی‌کنند زیاد است.

تایپ‌کردن: این پارامتر مشخص می‌کند آیا کاربر در طول فعالیت خود عبارتی را برای جستجو تایپ کرده است یا خیر. با توجه به این‌که ربات‌ها معمولاً برای پرس‌وجوهای خود از تایپ‌کردن استفاده نمی‌کنند بنابراین، این پارامتر می‌تواند نقش برجسته‌ای در تعیین انسان‌بودن کاربران ایفا می‌کند.

تعداد عامل‌های کاربر: اگر از یک آدرس شبکه بسته‌هایی با عامل‌های کاربر متفاوت دریافت شود، این پارامتر تعداد آن‌ها را مشخص خواهد کرد. البته این تعداد هم می‌تواند دلیلی بر ربات‌بودن کاربر موردنظر باشد و هم کاربرانی که با استفاده از فنّاوری NAT در تعامل با موتور جستجو هستند. البته با توجه به سایر پارامترها، مخصوصاً تعداد پرس‌وجوهای تکراری می‌توان به ماهیت این دسته از کاربران پی برد.

تعداد پرس‌وجوهای تکراری: ربات‌ها معمولاً پرس‌وجوهایی یکسان، و یا حداکثر پرس‌وجوهایی محدود از یک پایگاه داده معین را به تکرار برای موتور جستجو ارسال می‌کنند. درحالی‌که در سمت مقابل کاربران پرس‌وجوی پیشرفته‌تری را به کار می‌گیرند و به‌منظور حصول نتایج مطلوب خود، پرس‌وجوها را با تغییراتی ارسال خواهند نمود. لذا این پارامتر نیز تأثیر به سزایی در تعیین ربات در بر خواهد داشت.

رتبه نتایج انتخاب‌شده: نتایج ارائه‌شده توسط موتور جستجو براساس اعتبار و میزان ارتباط آن‌ها با پرس‌وجوی ارسال‌شده مرتب می‌شوند. لذا انسان‌ها معمولاً نتایج مطلوب خود را در بین نتایج صفحه نخست پیدا خواهند نمود و به‌ندرت نتایج سایر صفحات را بررسی می‌کنند. اما ربات‌هایی که برای افزایش مصنوعی رتبه یک صفحه مشخص به کارگرفته شده‌اند، به دنبال صفحه موردنظر خود در فهرست نتایج خواهند گشت و اگر این صفحه در نتایج اولیه از فهرست نتایج نباشد مکرراً صفحه‌ای با اولویت بیش از ۱۰ را انتخاب خواهند نمود.

تعداد فعالیت‌ها: این پارامتر مجموع تمام فعالیت‌های کاربر شامل پرس‌وجوها، انتخاب نتایج، انتخاب پیوندهای موجود در سایت و تایپ‌کردن را شامل می‌شود.

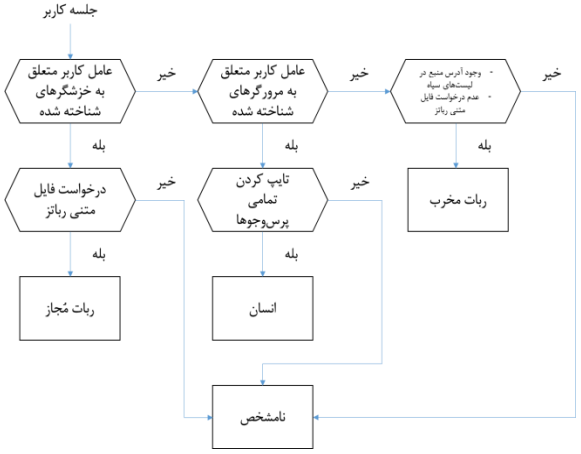
عامل کاربر: ربات‌هایی که به‌صورت خوش‌رفتار با موتور جستجو در تعامل هستند از همان ابتدا خود را در عامل کاربر ارسال‌شده به‌سوی موتور جستجو معرفی می‌کنند. البته این پارامتر به‌تنهایی باعث معاف‌شدن کاربر مورد نظر از بررسی‌های بیشتر نمی‌شود و فقط انتظار الگوریتم اتخاذشده را در ارائه رفتارهایی ربات‌گونه افزایش می‌دهد.

۴-۳- تولید داده آموزشی

برای تولید داده آموزشی از مجموعه داده ثبت رویداد تشریح‌شده فوق یک برنامه تحلیلگر داده‌های ثبت رویداد در موتور جستجو طراحی شد. این برنامه قادر است چهار برچسب ربات مخرب، ربات مجاز، انسان و نامشخص را به کاربران تخصیص دهد. قوانین طرح‌شده در برنامه تحلیلگر داده‌ها به شرح زیر است:

- جلسه‌های کاربری که دارای فیلد عامل کاربری متعلق به یک خزشگر شناخته شده است و در آن‌ها فایل متنی ربات درخواست شده با برچسب ربات مجاز مشخص می‌شوند.
- جلسه‌های کاربری با فیلد عامل کاربری متعلق به یک مرورگر شناخته‌شده که در طول فعالیت خود تمامی پرس‌وجوها را تایپ کرده‌اند به عنوان انسان برچسب زده می‌شوند.
- جلسه‌های کاربری با عامل کاربری ناشناخته که آدرس منبع آن‌ها در فهرست‌های سیاه آدرس شبکه وجود دارد و فایل متنی ربات را درخواست نکرده‌اند به عنوان ربات مخرب برچسب می‌خورند.
- سایر جلسه‌های کاربری عنوان نامشخص می‌گیرند.

بررسی ماهیت فیلد عامل کاربری از طریق آخرین نسخه پایگاه داده عامل‌های کاربری منتشرشده توسط [۲۱] انجام می‌شود. این پایگاه داده حاوی بیش از ۱۵۹ هزار عامل کاربری متعلق به خزشگرها و مرورگرهای شناخته‌شده می‌باشد. نحوه تعیین برچسب جلسات کاربران در شکل (۲) نشان داده شده است.



شکل (۲): تعیین برچسب کاربران در داده آموزشی

۴-۴- انتخاب بهترین صفات تشخیص

داده‌هایی که دارای ابعاد زیادی هستند علی‌رغم فرصت‌هایی که به‌وجود می‌آورند، چالش‌های محاسباتی را نیز ایجاد می‌کنند. یکی از مشکلات داده‌های با ابعاد زیاد این است که در برخی موارد، تمام صفات داده‌ها برای یافتن دانشی که در داده‌ها نهفته است، مهم و حیاتی نیستند. بعضی از صفات کاندید برای

یادگیری، نامربوط و زائد هستند و کارایی الگوریتم یادگیری را کاهش خواهند داد [۲۲]. بنابراین، انتخاب صفات مرتبط و ضروری در مرحله پیش‌پردازش از اهمیت به‌سزایی برخوردار است. فرآیند انتخاب صفت به معنای پیداکردن یک زیرمجموعه با حداقل اندازه ممکن، برای صفات است که برای هدف موردنظر اطلاعات لازم و کافی را دربرداشته باشد. بدیهی است که هدف تمام الگوریتم‌ها و روش‌های انتخاب صفت، همین زیرمجموعه است. روش‌های مختلف انتخاب صفت، تلاش می‌کنند تا از میان 2^n زیرمجموعه کاندید (n تعداد صفات)، بهترین زیرمجموعه را پیدا کنند.

برای این منظور، در این پژوهش از قابلیت نرم‌افزار وکا برای انتخاب صفات استفاده شده است. جستجو در فضای حالات در این پژوهش به روش اول بهترین[1] انجام می‌شود که به روش تپه‌نوردی حریصانه[2] سعی در انتخاب بهینه‌ترین زیرمجموعه ممکن از صفات دارد. ارزیاب زیرمجموعه بسته‌بندی[3] نیز برای ارزیابی زیرمجموعه‌ها استفاده شده است. دقت[4] زیرمجموعه‌های تعیین‌شده به کمک درخت تصمیم مورد ارزیابی قرار می‌گیرد و زیرمجموعه‌ای که در درخت تصمیم تشکیل‌شده بالاترین دقت را ارائه کند به عنوان زیرمجموعه بهینه انتخاب می‌شود.

نتیجه اجرای الگوریتم به این صورت است که از بین ۱۱ صفت ورودی، ۷۹ زیرمجموعه مورد ارزیابی قرار گرفته است و در نهایت، زیرمجموعه‌ای با ۵ صفت تعداد پرس‌وجو، تایپ‌کردن، نرخ ارسال درخواست، تعداد پرس‌وجوهای تکراری و نرخ انتخاب نتایج جستجو به عنوان زیرمجموعه بهینه انتخاب شده است.

۴-۵- ساخت مدل کلاس‌بندی

در طراحی مدل کلاس‌بندی، شناسایی ربات‌ها به صورت انفرادی و شبکه‌ای مورد توجه قرار گرفته است. در این روش ابتدا ربات‌هایی که در یک جلسه کاربری رفتار ناهنجاری را از خود نشان می‌دهند شناسایی شده و سپس در روش طراحی‌شده برای تشخیص ربات‌های شبکه‌ای، شناسایی شبکه‌هایی از ربات‌ها که برای افزایش رتبه صفحه مورد نظر خود کلمات کلیدی در یک زمینه خاص را به طور مکرر جستجو و نتیجه مورد نظر را انتخاب می‌کنند و همچنین شبکه‌هایی که کلمات کلیدی تکراری را برای ایجاد اختلال در سامانه پیشنهاددهنده موتور جستجو ارسال می‌کنند، مورد توجه قرار گرفته است.

تشخیص رفتار ناهنجار کاربران در یک جلسه کاربری با استفاده از درخت تصمیم صورت می‌گیرد. در زیر برخی از مزایای کلاس‌بندی براساس درخت تصمیم که آن را به روشی مناسب در این پژوهش تبدیل کرده است، بیان شده‌اند:

- درخت تصمیم، توانایی کار با داده‌های پیوسته و گسسته را دارد.
- درختان تصمیم، قادر به کار با داده‌های چندبعدی هستند.
- درخت تصمیم، از نواحی تصمیم‌گیری ساده استفاده می‌کند.
- مقایسه‌های غیرضروری در این ساختار حذف می‌شوند.
- از ویژگی‌های متفاوت برای نمونه‌های مختلف استفاده می‌کند.
- نیازی به تخمین تابع توزیع نیست.
- آماده‌سازی داده‌ها برای یک درخت تصمیم ساده یا غیرضروری است.
- گام‌های یادگیری و کلاس‌بندی از استنتاج درخت تصمیم ساده و سریع هستند.
- درخت تصمیم، یک مدل جعبه سفید است. توصیف شرایط در درختان تصمیم به آسانی و با منطق بولی امکان‌پذیر است. درحالی‌که برخی از الگوریتم‌ها به دلیل پیچیدگی در توصیف نتایج آن‌ها، مدل جعبه سیاه می‌باشند.
- نتایج ارائه‌شده در درخت، حالت شهودی داشته و عموماً توسط انسان به آسانی درک می‌شود.
- درخت‌های تصمیم، قادر به شناسایی تفاوت‌های زیرگروه‌ها می‌باشند [۲۳].

۴-۵-۱- تعیین اولویت صفات در درخت تصمیم

معیار انتخاب صفت یک فرآیند اکتشافی برای انتخاب ضابطه تقسیم است که به بهترین شکل اقدام به پارتیشن‌کردن داده از عناصر آموزشی برچسب خورده به کلاس‌های جداگانه می‌کند. همچنین، آن‌ها به‌عنوان قوانین تقسیم[5] نیز شناخته می‌شوند، زیرا تعیین می‌کنند که چگونه عناصر قرارگرفته در یک گره مشخص، تقسیم شوند. نمادهای به‌کاررفته در این بخش به شرح زیر هستند: D یک مجموعه داده از عناصر برچسب‌دار، m تعریف‌کننده تعداد کلاس‌های متمایز، $C_{i,D}$ مجموعه‌ای از عناصر از کلاس C_i در D باشد. |D| و $|C_{i,D}|$ به ترتیب دلالت بر تعداد عناصر در D و $C_{i,D}$ داشته باشند. در این مقاله از شاخص جینی به عنوان معیار انتخاب صفت، استفاده شده است. این شاخص ابتدا به اندازه‌گیری ناخالصی D به شکل زیر می‌پردازد:

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \qquad (۱)$$

که در آن، P_i احتمال این است که یک عنصر به کلاس C_i تعلق داشته باشد و توسط رابطه زیر تقریب زده می‌شود. مجموع بر روی m کلاس محاسبه می‌شود:

1- Best first
2- Greedy hillclimbing
3- Wrapper subset evaluator
4- Accuracy
5- Splitting rules

$$P_i = \frac{|C_{i,D}|}{|D|} \quad (۲)$$

شاخص جینی به‌عنوان یک تقسیم باینری برای هر صفت درنظر گرفته می‌شود. برای صفات با مقادیر پیوسته باید هر نقطه تقسیم ممکنه رسیدگی شود. در راه‌برد به‌کاررفته نقطه میانی بین هر جفت مقادیر مجاور (مرتب‌شده) می‌تواند به‌عنوان یک نقطه تقسیم ممکنه درنظر گرفته شود. نقطه‌ای که پایین‌ترین شاخص جینی برای یک صفت معین (با مقدار پیوسته) را دارد به‌عنوان نقطه تقسیم برای آن صفت انتخاب می‌شود. برای یک نقطه تقسیم ممکنه از A، D_1 مجموعه‌ای از عناصر در D است که شرط $A \leq split\_point$ و D_2 مجموعه‌ای از عناصر در D است که شرط $A > split_{point}$ را برآورده کند. به هنگام یک تقسیم باینری از رابطه زیر مبادرت به محاسبه وزن مجموع ناخالصی از هر پارتیشن به‌دست‌آمده می‌کنیم:

$$Gini_A(D) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2) \quad (۳)$$

کاهش در ناخالصی که می‌تواند به‌واسطه یک تقسیم باینری بر روی صفت A با مقدار گسسته یا پیوسته حاصل شود برابر است با:

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \quad (۴)$$

صفتی که کاهش در ناخالصی را به حداکثر برساند (یا برابر با پایین‌ترین شاخص جینی است) به‌عنوان صفت تقسیم‌کننده انتخاب می‌شود [۲۴]. جدول (۲) ترتیب صفات بررسی شده را به همراه ضرایب جینی و نقاط تقسیم آن‌ها را در اولین گام از درخت نمایش می‌دهد.

جدول (۲): ضرایب جینی صفات در درخت تصمیم

صفت	کاهش در ناخالصی	نقطه تقسیم
پرس‌وجوهای تکراری	۰/۲۳۹۵	۷/۵
تعداد پرس‌وجو	۰/۱۹۰۶	۴۸
نرخ ارسال درخواست	۰/۱۰۸۵	۵/۵
تایپ‌کردن	۰/۰۸۸	بله / خیر
نرخ انتخاب نتایج جستجو	۰/۰۵۰۲	صفر یا بیشتر

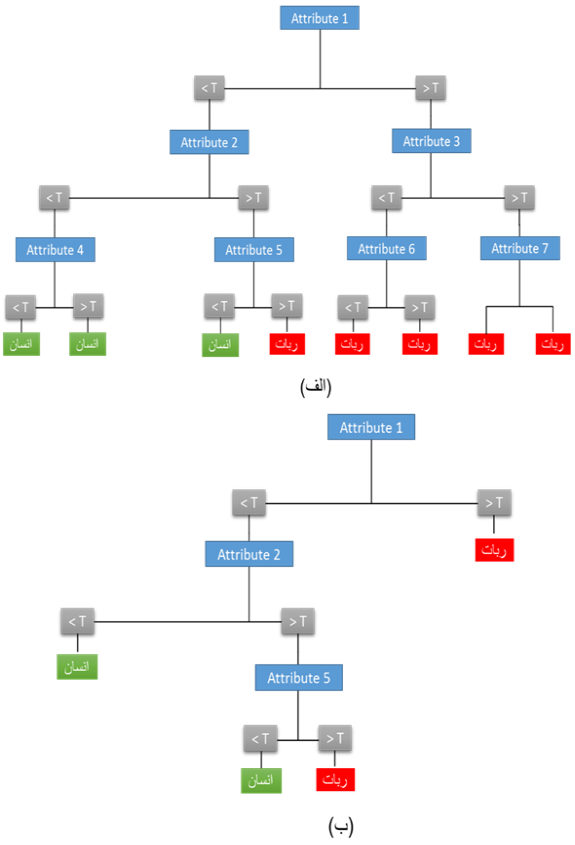
۴-۵-۲- هرس‌کردن درخت تصمیم

هرس درختان تمایل به ساخت درختان کوچک‌تر و با پیچیدگی کمتر دارد و از این‌رو، فهمیدن آن‌ها آسان‌تر می‌شود. معمولاً کلاس‌بندی داده‌های مستقل آزمون در چنین درخت‌هایی نسبت به درختان هرس‌نشده سریع‌تر انجام می‌شود. در این پژوهش، هرس‌کردن درخت تا رسیدن به درخت نهایی براساس دو شرط زیر صورت می‌گیرد:

۱- گره‌هایی که هیچ یک از داده‌های آموزشی از آن‌ها عبور نمی‌کنند، هرس می‌شوند.

۲- اگر دو گره برگ برادر دارای برچسب یکسانی باشند، هرس شده و والد آن‌ها به گره برگی با برچسب فرزندان خود تبدیل می‌شود. این فرآیند تا رسیدن به بالاترین سطح گره‌ها ادامه می‌یابد. شکل (۳) مثالی از این حالت را نمایش می‌دهد.

در انتهای این دو مرحله امکان اجرای درخت تصمیم بر روی تمام داده‌های ثبت رویداد ثبت‌شده از موتور جستجو فراهم خواهد شد.



شکل (۳): مثالی از درخت تصمیم (الف) پیش از هرس کردن (ب) پس از هرس‌کردن

۴-۵-۳- تشخیص شبکه‌های رباتی

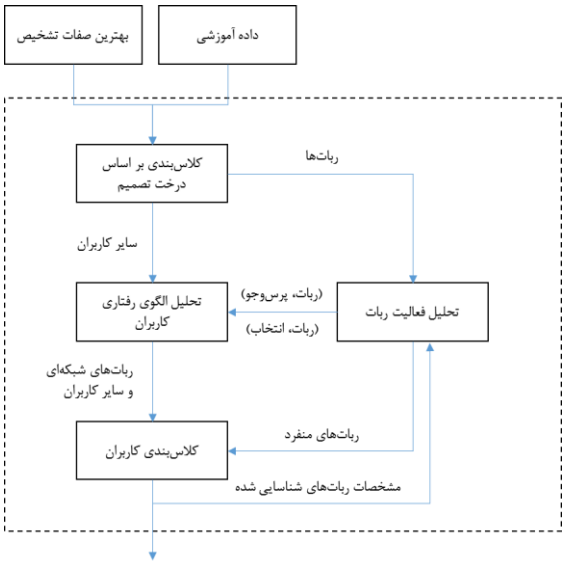
مهمترین اهداف ربات‌های جستجو را می‌توان شیوع پرس‌وجوهای خاص و افزایش مصنوعی رتبه برخی نتایج دانست. هدف اول واحدهای نمایه‌ساز و سامانه پیشنهاددهنده موتور جستجو را تحت تاثیر قرار می‌دهد و مورد دوم سامانه رتبه‌بند در موتور

جستجو را با اختلال همراه می‌کند. البته هردو این موارد منجر به ایجاد ازدحام و افزایش زمان پاسخگویی سامانه نیز خواهد شد. از طرفی ربات‌ها برای تحقق این اهداف ناچار به ارسال تعداد بسیار زیاد پرس‌وجوی تکراری و یا انتخاب مکرر نتایج موردنظر خود هستند. این ربات‌ها درصورتی‌که از یک ماشین واحد و یا تعداد محدودی از آن‌ها استفاده کنند، مجبور به ارسال حجم زیادی از درخواست‌ها از یک ماشین مشخص خواهند بود که این مساله شناسایی آن‌ها را ساده‌تر خواهد کرد. اما در مواردی، ربات‌ها برای جلوگیری از شناسایی‌شدن، ماشین‌های زیادی را قربانی اهداف خود کرده و با کمک آن‌ها، به نحوی‌که از هر ماشین تعداد درخواست‌های محدودی ارسال می‌شود، حمله گسترده‌ای را شکل می‌دهند. تمام روش‌های بررسی‌شده در بخش سوم (پیشینه تحقیق) و روش پیشنهادی که تا این قسمت تشریح شد تنها به بررسی فعالیت‌های یک کاربر و یا یک شبکه محلی (شبکه‌ای از ماشین‌ها با آدرس واحد)، بدون توجه به وجود ارتباط با سایر کاربران می‌پردازند و در مورد ماهیت آن‌ها تصمیم‌گیری می‌کنند. لذا این روش‌ها در تشخیص شبکه‌های رباتی مطرح‌شده ناتوان خواهند بود و نیاز به ارائه رویکرد نوینی در شناسایی این دسته از ربات‌ها وجود دارد.

همان‌طورکه گفته شد، ارسال پرس‌وجوهای تکراری زیاد و انتخاب مکرر نتایج مشخص، دو عامل دسترسی ربات‌ها به اهداف خود هستند. بنابراین، این دو عامل می‌توانند به عنوان کلید شناسایی این دسته از ربات‌ها مورد استفاده قرار گیرند.

ربات‌های شناسایی‌شده در درخت تصمیم، بخش دیگری از سامانه تشخیص ربات را با نام شناسایی الگوی شبکه‌های رباتی، فعال می‌کنند. ارتباط درخت تصمیم و بخش شناسایی الگوی شبکه‌های رباتی و در واقع معماری سامانه تشخیص ربات در شکل (۴) نشان داده شده است.

برای تشخیص ربات‌هایی که شیوع پرس‌وجوهایی در موتور جستجو را هدف خود قرار داده‌اند، در گام نخست پرس‌وجوهایی که توسط ربات‌های شناخته‌شده در درخت تصمیم، تکرار شده‌اند به عنوان پرس‌وجوهای مشکوک درنظر گرفته می‌شوند و در پایگاه داده مشخصی ذخیره می‌شوند. از طرف دیگر، کاربرانی که اقدام به ارسال هرکدام از این پرس‌وجوها می‌کنند نیز به عنوان کاربران مشکوک به ربات‌بودن درنظر گرفته می‌شوند. البته باید توجه داشت که در برخی رویدادهای فراگیر ممکن است دسته زیادی از کاربران پرس‌وجوهای مشابهی را ارسال کنند (مانند جستجو در مورد تیم فوتبال) لذا تنها ارسال پرس‌وجوهای مشابه نمی‌تواند دلیلی بر تشخیص یک شبکه باشد. اما تقریباً امکان این‌که کاربران انسانی الگوی یکسانی را در ارسال چندین پرس‌وجوی تکراری داشته باشند، وجود ندارد. بنابراین، با درنظرگرفتن این مساله می‌توان ربات‌هایی که پرس‌وجوهای خود را به صورت برنامه‌ریزی‌شده ارسال می‌کنند از انسان‌هایی که الگوی متفاوتی در ارسال پرس‌وجو دارند، تشخیص داد.



شکل (۴): معماری سامانه تشخیص ربات

در گام بعد، به هریک از پرس‌وجوهایی که در پایگاه داده ذخیره می‌شوند یک شناسه یکتا تعلق می‌گیرد و نیز شناسه پرس‌وجوهایی که یک کاربر اقدام به ارسال آن‌ها کرده است یک کد وضعیتی را برای کاربر متناظر شکل خواهد داد. در ادامه مثالی از چگونگی اختصاص کد وضعیتی به کاربران بررسی شده است.

$Q1 = c1$

$Q2 = c2$

$Q3 = c3$

$Q4 = c4$

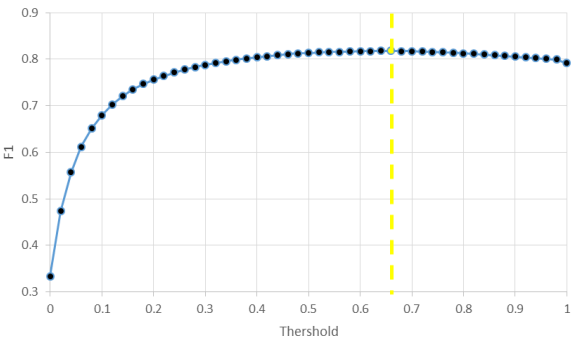
$U1\ Request\ Q1, Q3 \rightarrow SU1 = c1c3$

$U2\ Request\ Q2, Q4 \rightarrow SU2 = c2c4$

در مثال بالا، شناسه‌های c1، c2، c3 و c4 به ترتیب به پرس‌وجوهای Q1، Q2، Q3 و Q4 اختصاص داده شده است. کاربر U1 پرس‌وجوهای Q1 و Q3 و کاربر U2 پرس‌وجوهای Q2 و

Q4 را ارسال کرده‌اند. بنابراین، به ترتیب کدهای وضعیتی SU1 و SU2 به آن‌ها اختصاص داده شده است.

با اضافه‌شدن هر پرس‌وجو به فهرست پرس‌وجوهای مشکوک، گروهی با سرگروهی ربات متناظر شکل می‌گیرد. اگر پرس‌وجوهای چند ربات، یکسان و یا زیر مجموعه‌ای از یکدیگر بود در یک گروه قرار می‌گیرند. سپس با محاسبه کد وضعیتی برای هر کاربر، اگر شباهت پرس‌وجوهای کاربر با ربات سرگروه پرس‌وجو بیش از یک حد آستانه بود (رباتی‌که به‌واسطه آن پرس‌وجوی Q وارد فهرست پرس‌وجوهای مشکوک شده است)، این کاربر در گروه ربات موردنظر قرار می‌گیرد و اگر چندین ربات در گروه حضور داشته باشند الگوی رفتاری کاربر با آن‌ها نیز تطابق داده می‌شود. به این ترتیب، شبکه‌ای از ربات‌ها که اقدام به ارسال دسته مشخصی از پرس‌وجوها می‌کنند شناسایی خواهند شد. تعیین آستانه شباهت کاربران از مسائل مطرح در این روش است. برای این منظور، از مجموعه داده آموزشی استفاده شد و مقدار آستانه شباهت از صفر تا صد درصد افزایش یافت. برای هر کدام از مقادیر آستانه شباهت درنظر گرفته‌شده، معیار ارزیابی F1 محاسبه شد. این معیار ترکیبی از دقت و جامعیت در تشخیص ربات‌ها می‌باشد. نمودار شکل (۵) مقادیر مختلف محاسبه‌شده برای آستانه شباهت و بهترین مقدار این پارامتر را نشان می‌دهد. همان‌طورکه از نمودار مشخص است آستانه شباهت ۶۶ ٪ با بالاترین مقدار برای معیار F1 به عنوان مناسب‌ترین آستانه شباهت انتخاب شده است.



شکل (۵): تعیین آستانه شباهت برای تشخیص کاربران شبکه‌های رباتی

از نکات جالب توجه این روش این است که نه تنها کاربران شبکه‌های رباتی شناسایی می‌شوند؛ بلکه رفتار و هدف آن‌ها نیز شناسایی شده و با اضافه‌شدن کاربر جدید به هر گروه به سرعت تشخیص داده می‌شود.

همچنین روش مشابهی برای تشخیص ربات‌هایی که سعی در ایجاد اختلال در سامانه رتبه‌بندی موتور جستجو دارند اجرا خواهد شد. بدین ترتیب که این‌بار نتایج انتخاب‌شده توسط ربات‌ها به عنوان نتایج مشکوک درنظر گرفته شده و نگهداری می‌شود و کاربرانی که هریک از این نتایج را انتخاب کرده باشند در فهرست کاربران مشکوک قرار می‌گیرند. مشابه روش قبل به هر نتیجه موجود در پایگاه داده یک شناسه یکتا و به کاربران کد وضعیتی اختصاص داده می‌شود. بنابراین، در این روش شبکه‌های رباتی کلیک‌کننده و الگوی رفتاری آن‌ها شناسایی خواهد شد.

به‌کارگیری این روش در کنار درخت تصمیم موجود، امکان شناسایی ربات‌ها هم در قالب ماشین‌های واحد ناهنجار و هم در قالب شبکه‌های رباتی را فراهم می‌کند و قدرت قابل توجهی را به روش بوف می‌بخشد.

۴-۶- داده آزمون

به منظور تولید داده آزمون از روش به‌کار گرفته‌شده توسط Kang [۶] استفاده شد. در ادامه روش انتخاب‌شده برای انتخاب داده‌های آزمون توضیح داده خواهد شد.

به‌منظور انتخاب داده‌های آزمون که توزیع مناسبی بر روی انواع مختلف کاربران داشته باشد از یک روش تصادفی هوشمند استفاده شده است. در این روش ابتدا فلوچارتی براساس مقادیری اولیه که توسط تحقیقات گذشته در این حوزه تخمین زده شده است، ساخته شد. این ساختار براساس پنج پارامتر مختلف که در تحقیقات گذشته مقادیری برای آن‌ها تعریف شده است، تولید شد. سپس تمام داده‌ها توسط آن مورد بررسی قرار گرفته و هر کدام از کاربران به سمت گره‌های انتهایی هدایت می‌شوند. در انتها از هر گره انتهایی که دارای مقادیر می‌باشد کاربرانی به‌صورت تصادفی انتخاب می‌شوند. جدول (۳) پارامترهای اولیه مورد استفاده در این ساختار را نشان می‌دهد.

پس از انتخاب کاربران مناسب برای ساخت داده آزمون ابتدا تمامی پارامترهای تعریف‌شده برای کاربران محاسبه شده و علاوه‌بر آن‌ها محتوای پرس‌وجوها و نیز وجود آدرس فرستنده کاربران در فهرست‌های سیاه مختلف نیز مورد بررسی قرار گرفت. برای بررسی وجود آدرس فرستنده در فهرست‌های سیاه از برنامه موجود در [۲۵] استفاده شده است. این سایت آدرس درخواست‌شده را در ۸۰ پایگاه داده مختلف از فهرست‌های سیاه بررسی می‌کند و نتیجه هرکدام را بازمی‌گرداند. با اضافه‌شدن این دو پارامتر امکان تشخیص بهتری برای ماهیت کاربران به‌وجود خواهد آمد. در انتها برای هریک از کاربران برچسب‌هایی شامل انسان، ربات مخرب، ربات مجاز[1] و نامشخص درنظر گرفته شد.

1- Legal bot

جدول (۳): پارامترهای اولیه برای ساخت داده آزمون

پارامتر	انسان	ربات مخرب
تعداد درخواست در یک روز	کمتر از ۲۰۰ درخواست جستجو	بیش از ۲۰۰ درخواست جستجو
حداکثر نرخ درخواست	کمتر از ۷ درخواست در ۱۰ ثانیه	بیشتر یا مساوی ۷ درخواست در ۱۰ ثانیه
مدت‌زمان فعالیت	کمتر از ۱۲ ساعت در یک روز	بیشتر از ۱۲ ساعت در یک روز
نظم درخواست‌ها	بدون نظم	ارسال پرس‌وجوها در بازه‌های زمانی یکسان
نرخ کلیک کاربر	حداقل یک کلیک به ازای ۱۰ جستجو	بدون کلیک

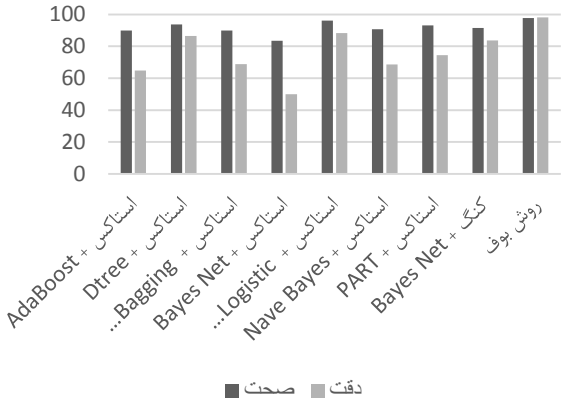
۵- ارزیابی عملکرد

در این بخش، مقایسه عملکرد روش بوف با روش‌های تشخیص ربات ارائه‌شده توسط Stokes و همکاران [۱۵] و Kang و همکاران [۶] انجام می‌گیرد. بررسی این روش‌ها در بخش سوم آورده شده است. در هر دو این روش‌ها، پیاده‌سازی با استفاده از نرم‌افزار داده‌کاوی وکا انجام گرفته است. بنابراین، در این پژوهش نیز پس از استخراج صفات تعریف‌شده در مقالات، پیاده‌سازی با استفاده از نرم‌افزار وکا صورت گرفته است. جدول (۴) نتایج حاصل از پیاده‌سازی تمام روش‌ها را نشان می‌دهد.

جدول (۴): مقایسه روش‌های مختلف تشخیص ربات در موتورهای جستجو

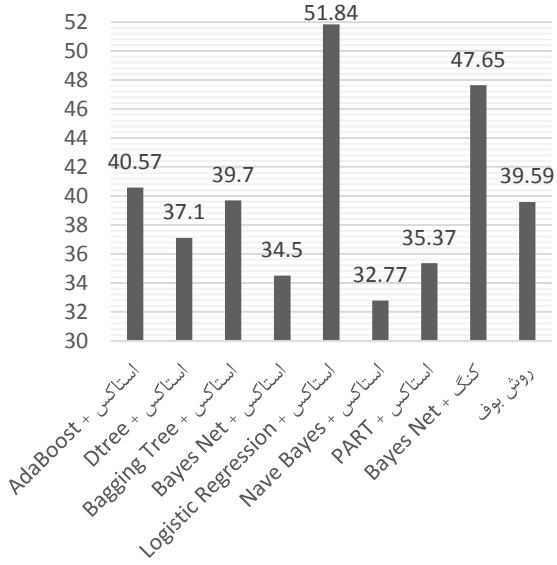
روش	کلاس‌بندی	مثبت حقیقی	منفی حقیقی	صحت	دقت	زمان اجرا
روش استاکس	AdaBoost	۹۰/۷	۸۶/۰	۹۰/۰	۶۴/۹	۴۰/۵۷
	DTree	۹۷/۷	۷۴/۴	۹۳/۸	۸۶/۴	۳۷/۱
	BaggingTrees	۹۳/۵	۷۲/۱	۹۰/۰	۶۸/۸	۳۹/۷
	Bayes Net	۸۰/۶	۹۷/۷	۸۳/۴	۵۰/۰	۳۴/۵
	Logistic Regression	۹۷/۷	۸۸/۴	۹۶/۱	۸۸/۳	۵۱/۸۴
	Nave Bayes	۹۲/۶	۸۱/۴	۹۰/۷	۶۸/۶	۳۲/۷۷
	PART	۹۴/۰	۸۸/۴	۹۳/۱	۷۴/۵	۳۵/۳۷
روش کنگ	Bayes Net	۹۶/۷	۷۰/۶	۹۱/۵	۸۳/۷	۴۷/۶۵
روش بوف	درخت تصمیم به همراه روش شناسایی الگو شبکه‌های رباتی	۹۷/۹	۹۶/۸	۹۷/۷	۹۸/۲	۳۹/۵۹

همان‌طورکه مشخص است سامانه ارائه‌شده در روش بوف عملکرد بسیار بهتری را در تشخیص ربات‌ها و انسان‌ها نشان می‌دهد. به نحوی‌که حداقل بهبود دقت ۹/۹ درصدی را نسبت به روش‌های استاکس و ۱۴/۵ درصدی را نسبت به روش کنگ نشان می‌دهد. البته این درصد زمانی تأثیر خود را بهتر نشان می‌دهد که برای تمام کاربران موتور جستجو درنظر گرفته شود. به‌طور مثال، برای کاربران تنها یک روز موتور جستجو این بهبود عملکرد، حداقل تصمیم‌گیری در مورد حدود ۲۲۳۰ کاربر انسانی را تحت تأثیر قرار می‌دهد. شکل (۶) تصویر مشخص‌تری را برای مقایسه روش‌های مختلف نشان می‌دهد.



شکل (۶): مقایسه عملکرد روش‌های مختلف تشخیص ربات

زمان اجرای این روش‌ها نیز در نمودار شکل (۷) مورد مقایسه قرار گرفته است.



شکل (۷): مقایسه زمان اجرای روش‌های مختلف تشخیص ربات

شاید در نگاه نخست زمان‌های ارائه‌شده توسط روش‌های

درخت تصمیم، شبکه Bayes، Nave Bayes و PART متعلق به استاکس و همکاران، اندکی بهتر از روش بوف به نظر آید؛ اما مقایسه عملکرد این روش‌ها نشان می‌دهد که به طور میانگین حدود ۷ ٪ صحت پایین‌تری را نسبت به روش بوف نشان می‌دهند. رقم قابل توجهی که تاثیر مختصر زمان افزایش یافته را از به خوبی از بین می‌برد.

۶- نتیجه‌گیری

این پژوهش تلاشی بود در جهت حفظ کارایی و البته جلوگیری از هدررفت منابع موتور جستجوی بومی یوز که برای نیل به آن، روشی جهت تشخیص ربات در پرس‌وجوهای موتور جستجو ارائه گردید. با بررسی‌هایی که بر روی کاربران موتورهای جستجو انجام گردید، کاربران در قالب سه گروه کلی دسته‌بندی شدند. این گروه‌ها شامل انسان‌ها، ربات‌ها و گروهی با عنوان نامشخص می‌شود. سپس تحقیقات سایر محققین فعال در این حوزه مورد ارزیابی قرار گرفت تا بتوان با بهره‌گیری از تلاش‌های آن‌ها به تشخیص هرچه بهتر و دقیق‌تر ماهیت کاربران موتورهای جستجو پرداخت. پس از آن درخت تصمیم به علت سادگی، سرعت و البته سربار کمی که به سامانه تحمیل می‌کرد به‌عنوان روشی مناسب جهت دسته‌بندی کاربران در گروه‌های مختلف انتخاب گردید.

در اولین گام جهت ساخت درخت تصمیم، نیاز به داده‌های آموزشی برچسب‌گذاری‌شده برای انتخاب صفات مناسب و نیز تشخیص اولویت آن‌ها به‌منظور قرارگیری در گره‌های درخت وجود داشت. لذا با توجه به پارامترهایی که برای تشخیص ربات‌ها تعریف شده بود، تصمیم به ساخت داده آموزشی از داده‌های ثبت رویداد مربوط به تعداد مشخصی از کاربران موتور جستجو یوز گرفته شد. با فراهم‌شدن داده آموزشی و به تبع آن مشخص‌شدن اولویت صفات در درخت تصمیم، نوبت به اجرای گام نهایی با نام هرس‌کردن درخت رسید. این گام به کمک تعریف قوانینی در درخت تصمیم انجام گردید. در ادامه روشی نوآورانه در شناسایی الگوی رفتاری شبکه‌های رباتی معرفی گردید. این روش پرس‌وجوهای ارسال‌شده توسط ربات‌ها و نتایج انتخاب‌شده توسط آن‌ها را که نیاز ربات‌ها برای تحقق اهدافشان می‌باشد، به عنوان کلیدی برای شناسایی آن‌ها درنظر گرفته و حتی ربات‌هایی که نرخ درخواست پایینی دارند را تشخیص خواهد داد. سپس روش ارائه‌شده با عنوان روش بوف نام‌گذاری شد. مجوز استفاده‌کردن و یا ردشدن صلاحیت روش بوف در گام ارزیابی بررسی می‌گردد؛ اما پیش از آن نیاز به یک مجموعه داده آزمون وجود دارد. ارزیابی کارایی روش بوف با مقایسه آن با روش‌های استاکس و کنگ صورت گرفت. این مهم به مدد معیارهای صحت و دقت و زمان اجرای الگوریتم انجام گردید. نتایج ارزیابی‌ها نشان می‌دهد که روش بوف قادر است در زمانی مناسب با دقتی بیش از ۹۷ ٪ ربات‌های فعال در موتور جستجو را تشخیص دهد. این رقم حداقل بهبود دقت ۹/۹ درصدی را نسبت به روش‌های بررسی‌شده در این حوزه نشان می‌دهد.

۷- مراجع

[1] A. ZareBidaki and F. KaveYazdi, "Big data management in search engines," In big data conference, 2015. (In Persian)

[2] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Inf. Sci. (NY), vol. 275, pp. 314–347, 2014.

[3] F. Yu, Y. Xie, and Q. Ke, "Sbotminer: large scale search bot detection," In Proceedings of the third ACM international conference on Web search and data mining, pp. 421–430, 2010.

[4] B. Kitts, J. Y. Zhang, G. Wu, W. Brandi, J. Beasley, K. Morrill, J. Ettedgui, S. Siddhartha, H. Yuan, and F. Gao, "Click Fraud Detection: Adversarial Pattern Recognition over 5 Years at Microsoft," In Real World Data Mining Applications, Springer, pp. 181–201, 2015.

[5] N. Sadagopan and J. Li, "Characterizing typical and atypical user sessions in clickstreams," In Proceedings of the 17th international conference on World Wide Web, pp. 885–894, 2008.

[6] H. Kang, K. Wang, D. Soukal, F. Behr, and Z. Zheng, "Large-scale bot detection for search engines," In Proceedings of the 19th international conference on World wide web, pp. 501–510, 2010.

[7] B. Kitts, J. Y. Zhang, A. Roux, and R. Mills, "Click Fraud Detection with Bot Signatures," in Intelligence and Security Informatics (ISI), 2013 IEEE International Conference, pp. 146–150, 2013.

[8] J. Zhang, Y. Xie, F. Yu, D. Soukal, and W. Lee, "Intention and Origination: An Inside Look at Large-Scale Bot Queries," in NDSS, 2013.

[9] J. P. John, F. Yu, Y. Xie, A. Krishnamurthy, and M. Abadi, "deSEO: Combating Search-Result Poisoning," in USENIX security symposium, 2011.

[10] S. Khattak, N. R. Ramay, K. R. Khan, A. Syed, and S. A. Khayam, "A taxonomy of botnet behavior, detection, and defense," Commun. Surv. Tutorials, IEEE, vol. 16, no. 2, pp. 898–924, 2014.

[11] N. Buzikashvili, "Sliding window technique for the web log analysis," in the 16th international conference on World Wide Web, pp. 1213–1214, 2007.

[12] Y. Zhang and A. Moffat, "Separating Human and Non-Human Web Queries," In the Web Information Seeking and Interaction Workshop, pp. 13–16, 2007.

[13] B. J. Jansen, A. Spink, and C. Blakely, "Defining a Session on Web Search Engines," vol. 58, no. 1998, pp. 862–871, 2007.

[14] N. Daswani and M. Stoppelman, "The Google click quality and security teams," The anatomy of Clickbot. A," In the First Workshop in Understanding Botnets, 2007.

[15] J. W. Stokes, G. Buehrer, K. Chellapilla, and J. C. Platt, "Classification of automated search traffic," In Weaving Services and People on the World Wide Web, Springer Berlin Heidelberg, pp. 3–26, 2009.

[16] O. Duskin and D. G. Feitelson, "Distinguishing humans from robots in web search logs: preliminary results using query rates and intervals," In Proceedings of the 2009 workshop on Web Search Click Data, pp. 15–19, 2009.

[17] A. Yasmin, M. C. Weigle, and M. L. Nelson, "Access Patterns for Robots and Humans in Web Archives," in 13th ACM/IEEE-CS joint conference on Digital libraries," ACM, 2013.

[18] M. Srivastava, A. K. Srivastava, R. Garg, and P. K. Mishra, "Comparative Analysis of Robot Detection Techniques on Web Server Log," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 9, pp. 186–189, 2015.

[19] W. Dong, X. Lei, Z. Hui, L. Hebing, Z. Hao, and S. Ting, "Web robot detection with semi-supervised learning method," in 3rd International Conference on Material, Mechanical and Manufacturing Engineering (IC3ME 2015), pp. 2123–2128, 2015.

[20] G. Buehrer, J. W. Stokes, and K. Chellapilla, "A large-scale study of automated web search traffic," In Proceedings of the 4th international workshop on Adversarial information retrieval on the web, pp. 1–8, 2008.

[21] "https://browscap.org/ua-lookup," 2016. [Online].

[22] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014.

[23] M. SanieiAbade, S. Mahmoodi, and M. Taherparvar, "practical data mining," 2nd ed. 2014. (In Persian)

[24] J. Han, M. Kamber, and J. Pei, "Data mining: comcepts and techniques," 3rd ed. 2011.

[25] "Blacklist Check." [Online]. Available: http://whatismyipaddress.com/blacklist-check.

Journal of Electronical & Cyber Defence
Vol. 5, No. 4, 2017, Serial No. 20

Detection of Anomaly Robots in the Search Engine Query

M. J. Sarvghad Moghaddam, M. Naghavi*, M. Ghayoori Sales

*Imam Hossein University

(Received: 16/08/2016, Accepted: 23/07/2017)

ABSTRACT

Search engines can be introduced as a best tool for managing, retrieving and extracting important information from a massive set of web data. These engines are scheduled to search the vast web environment and collect countless pages stored in every corner of the web. Search engines providers are always looking for improving the relationship between the results and reducing response times to users, but both of these can be influenced by the automated traffic sent by the bots. This article first defines bots and challenges of detecting them. Then, it provides a method named 'boof' for detecting Search robots. In 'the boof method', to achieve high accuracy in detecting anomaly robots, many different parameters are used to model the users' behavior. After determining the priority of parameters in detecting users, decision tree is made and attempted to categorize users into groups of humans, bots, legal bots and the unknown. Robots detected in the decision tree, enable another part of the robot detection system to identify robots even with low request rate. This is done by detecting the botnet behavior pattern . Evaluation of the proposed method on test data shows 97.7 percent accuracy in recognizing users that this improves the accuracy of at least 9,9 percent compared to the methods examined previously in this area. This is a significant digit that influences decision-making about 2230 users during each day.

Keywords: Search Engine, Search Robot, Log Analysis, Bot Detection, Decision Tree.

* Corresponding Author Email: mnaghavi@ihu.ac.ir