علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۱، ۱۳۹۵
صفحات ۳۲-۳۹
مقاله عادی

شناسائی عملی کلیک‌های هرز در وب با استفاده از الگوریتم‌های دسته‌بندی

مهدیه فلاح سجاد ظریف‌زاده

دانشکده برق و کامپیوتر، دانشگاه یزد، یزد، ایران

چکیده

امروزه اکثر سرویس‌های اینترنتی از بازخورد کاربران برای بهبود کیفیت سرویس‌دهی به آنان استفاده می‌نمایند. به عنوان مثال، موتورهای جستجو از اطلاعات کلیک کاربران به عنوان یک فاکتور مهم در فرآیند رتبه‌بندی نتایج جستجو بهره می‌برند. از همین‌رو، برخی وب سایت‌ها برای کسب رتبه بالاتر در بین مجموعه نتایج جستجو به انجام کلیک بر روی نتایج خود می‌پردازند. چون این کلیک‌ها توسط کاربران واقعی انجام نگرفته، اصطلاحاً به آنها کلیک‌های هرز گفته می‌شود. برای این منظور، وب سایت‌ها معمولاً از برنامه‌های نرم‌افزاری به نام "ربات‌ها" استفاده می‌کنند تا به صورت خودکار و توزیع‌شده به انجام این کار بپردازند. در این مقاله، روش جدیدی مبتنی بر دسته‌بندی نشست‌های کاربران جهت شناسائی کلیک‌های هرز به صورت سریع و کارآمد پیشنهاد می‌شود. ما در ابتدا نشست‌های کاربران را به صورت مجموعه‌ای از ویژگی‌ها مدل می‌کنیم و سپس با اعمال الگوریتم دسته‌بندی پیشنهادی، اقدام به شناسائی نشست‌های غیر نرمال و در نتیجه کلیک‌های هرز می‌نماییم. روش مطرح شده با لاگ واقعی یک موتور جستجوی فارسی مورد تحلیل قرار گرفته است. نتایج بررسی‌ها نشان می‌دهد که روش پیشنهادی می‌تواند کلیک‌های هرز را با دقتی بیش از ۹۶٪ تشخیص دهد که در مقایسه با کارهای قبلی در حدود ۵ درصد بهبود از خود نشان می‌دهد.

کلمات کلیدی: کلیک هرز، شناسائی ربات‌ها، ناهنجاری، یادگیری ماشین.

۱- مقدمه

امروزه موتورهای جستجو امکان دسترسی سریع، آسان و رایگان را به منابع عظیم اطلاعاتی موجود در سطح اینترنت برای کاربران فراهم می‌آورند. هنگامی که کاربر پرس‌وجوی خود را در موتور جستجو وارد می‌کند، آنها اسناد مرتبط با پرس‌وجوی کاربر را یافته و بر اساس فاکتورهای متعددی نظیر ویژگی‌های متنی [۱] و ساختار پیوندی بین صفحات [۲] رتبه‌بندی کرده و به کاربر نمایش می‌دهند. در دهۀ اخیر، موتورهای جستجو برای بهبود نتایج ارائه شده به کاربران، از کلیک‌های انجام شده روی مجموعه نتایج نیز به عنوان بازخورد مناسبی از سوی کاربران استفاده نموده و آن را در فرآیند رتبه‌بندی اسناد وارد می‌سازند. اکثر کاربران تنها به نتایج نخست (با رتبه بالاتر) توجه می‌نمایند [۳]. این امر می‌تواند موجب سوءاستفاده از موتورهای جستجو و دست‌کاری صفحۀ نتایج به منظور بالا بردن رتبۀ برخی صفحات خاص و یا احیاناً خرابکاری شود [۴]. حمله‌کنندگان با استخدام مجموعه‌ای از افراد و یا با استفاده از ربات‌ها (برنامه‌های نرم‌افزاری که به صورت خودکار به ارسال پرس‌وجو و یا انجام کلیک روی لینک‌ها می‌پردازند) به صورت توزیع شده، به انجام حملات مختلف دست می‌زنند. چون این کلیک‌ها توسط کابران واقعی انجام نمی‌شود، اصطلاحاً به آنها "کلیک‌های هرز" گفته می‌شود.

مسئله شناسائی و تفکیک ترافیک تولید شده توسط ربات‌ها از ترافیک کاربران واقعی و نرمال برای موتورهای جستجو بسیار حائز اهمیت است، زیرا وجود ترافیک‌های غیر نرمال علاوه بر تغییر در رتبه‌بندی نتایج جستجو می‌تواند با مصرف پهنای باند موتور جستجو، افزایش زمان پاسخگویی به کاربران واقعی و تاثیر منفی روی تصمیم‌گیری‌هایی که براساس سابقه و بازخورد کاربران گرفته می‌شود [۵]، به موتور جستجو صدمه بزنند.

از سوی دیگر، درآمد اصلی سرویس‌های رایگانی نظیر موتورهای جستجو از سیستم تبلیغات آنلاین آنها می‌باشد. افزایش روز افزون کاربران اینترنتی نیز به رونق این کسب و کار کمک شایانی نموده به نحوی که درآمد حاصل از تبلیغات آنلاین در سال ۲۰۱۵ به ۵۹/۶ میلیارد دلار رسیده است که این مقدار نیست به سال قبل خود، بیشتر از ۱۵٪ رشد داشته است [۶]. از این منظر، موتورهای جستجو به عنوان شبکه‌های تبلیغاتی به نمایش لینک‌های تبلیغاتی اقدام می‌کنند.

آنها از دو روش ۱) نمایش تبلیغات مرتبط با پرس‌وجوی کاربر در کنار نتایج جستجو و ۲) نمایش تبلیغات گرافیکی در وب‌سایت‌های مرتبط با محتوای تبلیغ (یا اصطلاحاً منتشرکنندگان) بهره می‌گیرند. شبکه‌های تبلیغاتی معمولاً بر مبنای مدل "پرداخت به ازای هر کلیک" [۷] فعالیت می‌نمایند یعنی هر زمان روی یک تبلیغ کلیک شود، مبلغی از شارژ تبلیغ‌کننده کسر می‌گردد. در حالت دوم، موتورهای جستجو بخشی از درآمد هر کلیک را به منتشرکننده‌ای که کلیک از طریق آن انجام گرفته است، می‌دهند.

در سیستم تبلیغات آنلاین، کلیک‌های هرز می‌تواند توسط منتشرکنندگان و یا تبلیغ‌کنندگان انجام گیرد. منتشرکنندگان متقلب با انگیزهٔ کسب درآمد بیشتر و همچنین تبلیغ‌کنندگان با هدف تمام کردن بودجه تبلیغات رقیب خود به این کار مبادرت می‌ورزند. گرچه با وجود این کلیک‌ها، موتورهای جستجو باز هم به ازای هر کلیک، درآمد خود را کسب می‌کنند اما بی‌توجهی به این مسئله در بلندمدت موجب از بین رفتن اعتبار آنها نزد تبلیغ‌کنندگان می‌شود. بنابراین، شناسائی این نوع از کلیک‌ها و کسر نکردن شارژ تبلیغ‌کنندگان به ازای آنها بسیار مهم می‌باشد.

هرچند مسئله کلیک‌های هرز در سایر سیستم‌های آنلاین نظیر سیستم‌های توصیه‌گر و شبکه‌های اجتماعی که از اطلاعات کلیک کاربران در تصمیم‌گیری‌های خود استفاده می‌کنند، وجود دارد، اما در مورد موتورهای جستجو به دلیل حجم زیاد کاربران و ترافیک بسیار بالای آنها از اهمیت بسزایی برخوردار است. لذا ما در این مقاله، صرفاً بر روی مسئله شناسائی کلیک‌های هرز در موتورهای جستجو تمرکز می‌نماییم.

کلیک‌های هرز در موتورهای جستجو براساس اهدافشان به دو نوع مختلف تقسیم‌بندی می‌شوند: ۱) کلیک‌های هرز روی نتایج اصلی موتورهای جستجو با هدف افزایش رتبهٔ یک وب‌سایت در صفحهٔ نتایج و ۲) کلیک‌های هرز روی لینک‌های تبلیغاتی موجود در صفحهٔ نتایج با هدف تمام کردن بودجه یک تبلیغ‌کنندهٔ خاص. موتورهای جستجو باید صرف نظر از نوع کلیک‌های هرز، آنها را شناسائی و از ترافیک کاربران نرمال تمایز دهند.

در گذشته، معمولاً حمله‌کنندگان از تعداد ثابتی آدرس IP برای تولید ترافیک غیر نرمال استفاده می‌کردند، لذا شناسائی آنها نسبتاً ساده بود. اما به تدریج ابزارهای آنها پیشرفت کرد به نحوی که اکثر حملات امروزی به صورت کاملاً خودکار و توزیع شده توسط شبکه‌ای از ربات‌ها و بدافزارها انجام می‌گیرد [۸-۱۱]، بنابراین شناسائی آنها بسیار دشوار و پیچیده شده است. به عنوان مثال، در [۱۰] بدافزاری تشریح شده است که توانست روی بیش از ۴ میلیون کامپیوتر بنشیند و با تغییر آدرس "سرور نام میزبان"[1] آنها به نمایش لینک‌های تبلیغاتی و انجام کلیک‌های هرز روی آنها بپردازد. این بدافزار به مدت ۴ سال ناشناخته ماند و ۱۴ میلیون دلار برای صاحبان خود به ارمغان آورد. لذا، به دلیل اهمیت موضوع، محققان زیادی از سراسر جهان به موضوع شناسائی ترافیک‌های غیر نرمال و تفکیک آنها از ترافیک کاربران واقعی روی آوردند.

در این مقاله، ما روش جدیدی مبتنی بر دسته‌بندی برای مقابله با کلیک‌های هرز، ارائه می‌کنیم. روش پیشنهادی شامل ۱) مدل‌سازی نشست‌های کاربران، ۲) دسته‌بندی نشست‌های مدل شده به دو دستهٔ "نرمال" و "غیر نرمال" و ۳) به روزرسانی و تقویت مجموعه آموزشی می‌باشد. در این پژوهش، بر خلاف کارهای قبلی، ما با ترکیب سه سطح از ویژگی‌ها (نشست، کاربر و آدرس IP) سعی نمودیم تمامی جنبه‌های رفتاری غیر نرمال را پوشش داده و از آنها برای تفکیک کاربران انسانی و ربات‌ها بهره بگیریم. یکی دیگر از نقاط ممتاز این پژوهش، راهکار "تقویت مجموعه آموزشی" می‌باشد که به کمک آن می‌توان ضمن ایجاد پویایی در مجموعه داده‌های آموزشی، دقت دسته‌بندی را نیز افزایش داد.

ساختار مقاله در ادامه به صورت زیر می‌باشد: ابتدا کارهای مرتبط با موضوع پژوهش را در بخش دوم، مرور می‌نماییم. در بخش سوم، سیستم پیشنهادی که شامل مجموعه ویژگی‌های استفاده شده و الگوریتم دسته‌بندی پیشنهاد شده جهت شناسائی کلیک‌های هرز می‌باشد را معرفی و تشریح می‌کنیم. در بخش چهارم به ارزیابی روش پیشنهاد شده و مقایسه آن با کارهای قبلی می‌پردازیم و نهایتاً در بخش پنجم به جمع‌بندی و ارائه پیشنهاداتی برای ادامه کار خواهیم پرداخت.

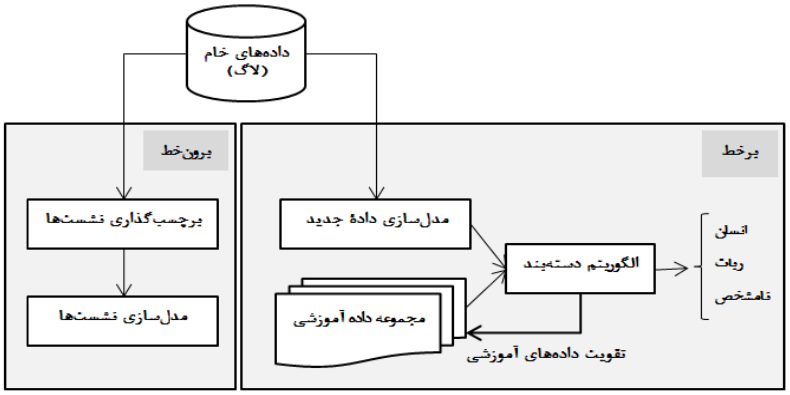
۲- کارهای مرتبط

در یک دههٔ گذشته، شناسائی کلیک‌های هرز در شبکه‌های تبلیغاتی بسیار مورد توجه قرار گرفته است. استون- گراس و همکارانش [۱۲] با استفاده از لاگ جستجوی یک تبادلگر تبلیغ[2] اینترنتی به بررسی طیف گسترده‌ای از مشخصات فعالیت کاربران نامتعبر شامل رفتارهای مرتبط با کلیک‌های هرز پرداختند. روش دیگری که توسط دیو و همکارانش [۱۳] پیشنهاد شد، ابتدا کاربرانی را که بیشترین درآمد برای هر منتشرکننده داشته‌اند، پیدا می‌کند و سپس با مقایسهٔ درآمد آنها با درآمد تولید شده از کاربران متناظر با منتشرکنندگان درستکار، منتشرکنندگان متقلب را شناسائی می‌نماید. در [۱۴]، مجموعه‌ای از محققان با کمک تکنیک‌های مختلف داده‌کاوی، الگوهای مشترک در بین کلیک‌های هرز را تشخیص می‌دهند. در پژوهشی مشابه، کیت و همکارانش [۱۵]، با تکیه بر تجربه ۵ ساله خود در شبکه تبلیغات مایکروسافت، چالش‌های توسعه یک سیستم داده‌کاوی جهت شناسائی کلیک‌های هرز را بیان و پارامترهای مختلف در طراحی چنین سیستمی را تشریح می‌کنند. تبلیغات بلوف [۱۶] و ایجاد امضا برای بات‌ها [۱۷] از دیگر روش‌هایی هستند که با تشخیص کاربران معتبر، به مقابله با حملات کلیکی در شبکه‌های تبلیغاتی می‌پردازند. کارهای معرفی شده تاکنون، عمدتاً به مسئله شناسائی کلیک‌های هرز در شبکه‌های تبلیغاتی پرداخته‌اند، لذا مستقیماً به موضوع این مقاله مربوط نیستند. در ادامه تعدادی از روش‌هایی را که صرفاً به شناسائی کلیک‌های هرز در موتورهای جستجو (کلیک‌های هرز روی لینک‌های تبلیغاتی و نتایج معمولی) پرداخته‌اند و ما نیز در این مقاله از برخی ایده‌های آنها استفاده نموده‌ایم، مرور می‌کنیم.

یو و همکارانش [۱۸]، سیستمی ارائه دادند که می‌تواند ترافیک‌های تولید شده توسط ربات‌ها را شناسائی نماید. آنها گروه‌هایی را از کاربران که حداقل یک پرس‌وجو/کلیک مشترک دارند، یافته و به کمک محاسبات ماتریسی به شباهت‌سنجی تاریخچه فعالیت‌های آنها می‌پردازند. این سیستم قادر به شناسائی ترافیک‌های غیر نرمالی که با نرخ پایین اما به صورت توزیع‌شده توسط شبکه‌ای از ربات‌ها ارسال می‌شود، می‌باشد. در [۱۹]، از تکنیک‌های یادگیری ماشین برای شناسائی کلیک‌های هرز استفاده می‌شود. به این ترتیب که ابتدا با به کارگیری کدهای کپچا و تعدادی روش شهودی، حجم زیادی داده آموزشی برچسب‌دار تولید و سپس با توسعه یک الگوریتم نیمه نظارتی از قدرت داده‌های فاقد برچسب برای بهبود کارایی دسته‌بندی‌کننده استفاده می‌شود.

روش‌های جدیدتر با بررسی فعالیت کاربران در سطح نشست به شناسائی کلیک‌های هرز می‌پردازند. محققان در [۲۰]، هر نشست از کاربران را با کمک زنجیره مارکوف به صورت دنباله‌ای از جفت‌های (نوع فعالیت کاربر، شماره صفحه) مدل کرده و فاصلهٔ هر نشست از میانگین نشست‌ها را محاسبه می‌نمایند. فاصلهٔ زیاد بیانگر نشست‌های غیر نرمال بوده و تمام کلیک‌های انجام شده در آن به عنوان کلیک هرز در نظر گرفته می‌شود. در [۲۱] نیز نشست کاربران مدل می‌شود اما با توالی‌های سه تایی از (نوع فعالیت کاربر، هدف فعالیت و اختلاف زمانی هر فعالیت نسبت به فعالیت قبلی خود). بعد از مدل‌سازی نشست‌های کاربران، دو الگوریتم انتشار گراف دوبخشی "کاربر- نشست" و "نشست- الگو" برای شناسائی کلیک‌های هرز پیشنهاد می‌شود.

روش‌هایی که ذکر شد هر یک به جنبه خاصی از رفتارهای غیر نرمال پرداخته و تنها قادر به شناسائی حملات انجام شده در آن دسته هستند.



شکل ۱- شمای کلی از سیستم پیشنهادی

در این مقاله، ما جنبه‌های مختلفی از رفتار ترافیک‌های غیر نرمال را به صورت مجموعه‌ای از ویژگی‌ها در سه سطح نشست، کاربر و IP با یکدیگر ترکیب نموده و با کمک تکنیک دسته‌بندی پیشنهاد شده که گونه‌ای تغییر یافته از الگوریتم K-نزدیک‌ترین همسایه می‌باشد، به شناسائی کلیک‌های هرز می‌پردازیم. همچنین بیشتر روش‌های قبلی به صورت آفلاین کار می‌کردند اما سیستم مطرح شده در این مقاله، قابلیت به کارگیری برخط را نیز دارا می‌باشد گرچه این قابلیت در اینجا مورد ارزیابی قرار نگرفته است.

۳- روش پیشنهادی

همانطور که در بخش قبل اشاره شد، روش‌هایی که اخیراً در زمینه شناسایی کلیک‌های هرز مطرح شده‌اند، عموماً به شناسایی نشست‌های غیر نرمال متکی هستند. سیستمی که ما نیز در این مقاله پیشنهاد می‌دهیم در این دسته قرار می‌گیرد، با این تفاوت که با افزودن ویژگی‌هایی در سطح کاربر و آدرس IP سعی می‌کنیم دامنهٔ تشخیص خود را گسترده‌تر نموده تا بتوانیم رفتارهای مختلف غیر نرمال و ترافیک رباتی را با کمک این سیستم شناسایی نماییم.

شمای کلی از سیستم پیشنهاد شده در شکل ۱ نشان داده شده است. ما در این مقاله، از تکنیک دسته‌بندی نشست‌های کاربران بهره می‌گیریم. بنابراین روش مطرح شده شامل دو بخش آموزش و آزمایش (یا به عبارت دیگر برون‌خط و برخط) می‌باشد. در فاز آموزش، نخست به معرفی ویژگی‌ها و مدل‌سازی داده‌ها خواهیم پرداخت. سپس، چالش‌های موجود جهت تولید مجموعه داده آموزشی برچسب‌دار را مطرح نموده و با مرور تعدادی از روش‌های عنوان شده، روش مناسب خود را انتخاب می‌نماییم. در فاز آزمایش، الگوریتم دسته‌بندی پیشنهادی که گونه‌ای تغییر یافته از الگوریتم K- نزدیک‌ترین همسایه می‌باشد را معرفی می‌کنیم. در انتها، تکنیکی جهت تقویت مجموعه داده آموزشی و افزایش دقت دسته‌بندی ارائه می‌دهیم.

۳-۱- مدل‌سازی داده‌ها

داده‌های استفاده شده در این پژوهش، لاگ جستجوی یکی از پربازدیدترین موتورهای جستجوی فارسی (به آدرس parsijoo.ir) می‌باشد. هر رکورد از این مجموعه لاگ، یک درخواست از سوی کاربر است که حاوی اطلاعات زمانی، پرس‌وجوی ارسال شده، لینک کلیک شده، آدرس IP و شناسه کاربر می‌باشد. درخواست‌ها به ترتیب زمان واقعی‌شان پیمایش شده و به ازای هر درخواست مجموعه‌ای از ویژگی‌ها محاسبه شده و پس از دسته‌بندی در مورد نرمال یا غیر نرمال بودن آن تصمیم‌گیری می‌شود.

مجموعه ویژگی‌های استخراج شده از داده به سه سطح تقسیم می‌شوند: ۱) ویژگی‌های رفتاری سطح نشست ۲) ویژگی‌های رفتاری سطح کاربر و ۳) ویژگی‌های رفتاری سطح IP که به ترتیب به معرفی هر یک می‌پردازیم.

۳-۱-۱- ویژگی‌های سطح نشست

برای محاسبه ویژگی‌های سطح نشست، ما فعالیت‌های کاربر در هر نشست را مورد بررسی قرار می‌دهیم. زمانی‌که یک کاربر وارد موتور جستجو می‌شود، یک شناسه یکتا به عنوان "شناسه نشست" به او تخصیص داده می‌شود. این شناسه بعد از گذشت یک محدودیت زمانی (معمولاً ۳۰ دقیقه) از فعال نبودن کاربر منقضی می‌شود و وقتی آن کاربر مجدداً به سایت برگشت، یک شناسه نشست جدید به او داده می‌شود. هر کاربر ممکن است در طول نشست فعالیت‌های مختلفی در سیستم انجام دهد: به ارسال پرس‌وجو بپردازد، صفحه نتایج را مرور نماید، روی لینک‌های نتایج کلیک کند، روی یک صفحه خاصی از مجموعه صفحات نتایج کلیک کند، پرس‌وجوی خود را اصلاح نماید و غیره. در این مقاله ما تنها سه نوع فعالیت را در نظر می‌گیریم:

- Q_i: ارسال یک پرس‌وجو (که i به معنای پرس‌وجوهای مختلف می‌باشد، برای مثال، Q_1 بیانگر یک پرس‌وجو است و Q_2 بیانگر یک پرس‌وجوی دیگر).
- W_i: کلیک روی لینک نتایج جستجو یا کلیک روی لینک‌های موجود در صفحه نخست (که i به معنای لینک‌های متفاوت می‌باشد).
- N: کلیک روی شماره صفحات مختلف از مجموعه صفحات نتایج. این فعالیت می‌تواند شامل کلیک روی دکمه "صفحه بعد"، "صفحه قبل" و یا یک شماره صفحه خاص باشد.

برای بررسی رفتار کاربر در هر نشست، مجموعه ویژگی‌های زیر از نشست جاری او استخراج می‌شود:

- احتمال مارکوف دنباله فعالیت‌های کاربر در نشست: این ویژگی در [۲۰] مطرح شده و مقدار آن پس از مدل‌سازی فعالیت‌های کاربر به صورت زنجیره مارکوف، از حاصل ضرب احتمال انتقال از یک وضعیت به وضعیت بعدی به دست می‌آید.
- تعداد کل فعالیت‌های کاربر در نشست برحسب (Q، W، N)
- تعداد کل کلیک‌های انجام شده روی نتایج وب
- تعداد کل پرس‌وجوهای ارسال شده
- تعداد کل کلیک‌های انجام شده روی شماره صفحات دیگر
- نسبت لینک‌های یکتای کلیک شده به کل لینک‌های کلیک شده
- نسبت دامنه‌های کلیک شده به کل لینک‌های کلیک‌شده

- نسبت پرس‌وجوهای یکتای ارسال شده به کل پرس‌وجوهای ارسال شده
- نسبت مجموع فعالیت‌های یکتای انجام شده به کل فعالیت‌ها
- نسبت پرس‌وجوهای ارسال شده به مدت زمان فعالیت نشست
- نسبت کلیک‌های انجام شده به مدت زمان فعالیت نشست
- نسبت کلیک‌های روی شماره صفحات دیگر به مدت زمان فعالیت نشست

هفت ویژگی آخر برای اولین بار در این مقاله پیشنهاد شده‌اند و مابقی در [۲۰، ۲۲] به کار رفته‌اند. مقادیر تمام ویژگی‌های فوق و همین‌طور ویژگی‌هایی که در ادامه خواهند آمد به بازه [۱-۰] نرمال‌سازی می‌شوند.

۳-۱-۲- ویژگی‌های سطح کاربر

ویژگی‌های مطرح شده در سطح نشست، از نشست جاری کاربر محاسبه می‌گردند اما هر کاربر ممکن است تاکنون نشست‌های متعددی داشته باشد. بنابراین می‌توان تمام تاریخچه نشست‌های کاربر را در قالب ویژگی‌های سطح کاربر لحاظ کرد. ویژگی‌های مطرح شده در سطح کاربر، عیناً مانند ویژگی‌های سطح نشست هستند با این تفاوت که از میانگین کلیه نشست‌های کاربر محاسبه می‌گردند.

۳-۱-۳- ویژگی‌های سطح IP

بسیاری از ربات‌ها قادر به اجرای کدهای جاوا اسکریپت نیستند و یا کوکی آنها غیر فعال است. بنابراین به ازای هر درخواست که از جانب آنها به موتور جستجو می‌آید یک شناسه کاربری جدید به آنها تخصیص می‌یابد، در نتیجه هر نشست از آنها تنها شامل یک فعالیت است. بنابراین ویژگی‌های سطح نشست و سطح کاربر به تنهایی کافی نیستند بلکه نیاز به وجود ویژگی‌هایی در سطح IP نیز می‌باشد. ویژگی‌هایی که ما برای هر IP در نظر گرفتیم عبارتند از:

- آدرس IP اینترانت یا اینترنت (0/1): کاربرانی که از سازمان‌ها و ادارات داخل کشوری هستند دارای آدرس IP اینترانت می‌باشند.
- تعداد کل فعالیت‌های IP
- تعداد پرس‌وجوهای ارسال شده
- تعداد کل کلیک‌های انجام شده روی نتایج جستجو
- تعداد کل کلیک‌های انجام شده روی شماره صفحات
- نسبت لینک های یکتای کلیک شده به کل لینک‌های کلیک شده
- نسبت دامنه‌های یکتای کلیک شده به کل لینک‌های کلیک شده
- نسبت کوکی‌های تخصیص یافته به کل فعالیت‌های هر IP

هر نشست از کاربر ممکن است با یک یا چند آدرس IP همراه باشد بنابراین ویژگی‌های فوق به صورت میانگین روی تمامی آدرس‌های آن نشست محاسبه می‌گردد.

۳-۲- تولید مجموعه داده آموزشی اولیه

حجم بالای لاگ موتورهای جستجو، برچسب‌گذاری دستی آنها به عنوان "انسان/ربات" جهت تولید مجموعه داده آموزشی را تقریباً ناممکن می‌سازد. در برخی از سرویس‌های اینترنتی مانند ایمیل یا بانکداری الکترونیک می‌توان برای دسترسی به خدمات از مکانیزم کپچا جهت احراز هویت کاربران و تمایز آنها از ربات‌ها استفاده نمود. به این ترتیب تنها کاربران واقعی امکان دسترسی به سرویس را خواهند داشت و فعالیت ربات‌ها پشت کپچا متوقف می‌شود. اما این روش برای موتورهای جستجو کاربردی نیست زیرا هدف موتورهای جستجو، سرعت و سهولت در ارائه خدمات با کمترین فعالیت اضافه از سوی کاربر می‌باشد. با این وجود، در [۲۳] ایدهٔ استفاده از کدهای کپچای کلیک‌پذیر مطرح شده است. با به کارگیری این نوع از کدهای کپچا کاربران کمتر به زحمت می‌افتند و می‌توانند با سرعت و دقت آن را پشت سر بگذارند. اما همان‌طور که گفته شد، ایده نمایش کپچا به تمام کاربران موتور جستجو ایده خوبی نیست. کنگ و همکاران [۱۹]، برای تولید مجموعه داده آموزشی، پیشنهاد نمایش کپچا تنها به بخش کوچکی از کاربران را مطرح کردند. در این روش، تنها به کاربرانی که از آستانه‌های شهودی تعریف شده برای چند پارامتر ساده عبور می‌کنند، کدهای کپچا نمایش داده می‌شود. نویسندگان مقاله ادعا کردند که با این روش، تنها از ۱٪ از کاربران خواسته شده که به کپچا پاسخ دهند که اکثرا نیز به آن جواب نداده‌اند که این به معنای برنامه‌های رباتی می‌باشد که قادر به حل کپچا نیستند.

ما نیز مشابه [۱۹]، با استفاده از تعدادی روش شهودی ساده نظیر حجم فعالیت‌های کاربر در یک بازه زمانی کوتاه و لیست سیاهی از آدرس‌های IP، مجموعه اولیه‌ای از نشست‌های غیر نرمال تولید نمودیم. در این حالت، اگر رفتار یک کاربر از آستانه‌های تعریف شده فراتر رفت، کد کپچا به سوی او ارسال می‌گردد. کاربر ممکن است به کد کپچا پاسخ ندهد، به اشتباه پاسخ دهد و یا به درستی پاسخ دهد. در دو وضعیت نخست، ما نشست کاربر را به عنوان "مثبت" یا "ربات" برچست می‌زنیم. به این ترتیب، مجموعه‌ای از نشست‌های غیر نرمال تولید می‌شود. در نهایت، ویژگی‌های سطح نشست، کاربر و آدرس IP به ازای آنها محاسبه شده (طبق بخش ۳-۱) و از آنها به عنوان مجموعه داده آموزشی اولیه در فرآیند دسته‌بندی استفاده می‌گردد. اما با روشی که در ادامه خواهیم گفت به تدیج، به بهبود کیفیت مجموعه آموزشی کمک کرده و می‌توانیم با دقت بالاتری نسبت به دسته‌بندی نشست‌ها اقدام نماییم.

برای تولید مجموعه داده گفته شده، ما از دو هفته لاگ جستجو و کلیک استفاده نمودیم (از ۹۴/۰۹/۱۰ تا ۹۴/۰۹/۱۶) که شامل بیش از دو میلیون و سیصد هزار درخواست (پرس‌وجوها و کلیک‌ها) و بیش از نهصد و سی هزار نشست یکتا بود.

۳-۳- الگوریتم دسته‌بندی

در روش دسته‌بندی مطرح شده ما از الگوریتم K- نزدیک‌ترین همسایه[۳] (KNN) [۲۴]، به عنوان الگوریتم پایه استفاده کرده و سپس بنا به ضرورت، تغییراتی در آن لحاظ نمودیم. الگوریتم K- نزدیک‌ترین همسایه یک روش غیر پارامتری می‌باشد که به دلیل سادگی، سرعت و کارایی در بسیاری از مسائل دسته‌بندی و رگرسیون به عنوان مناسب‌ترین روش مورد استفاده قرار می‌گیرد. این الگوریتم برای دسته‌بندی یک دادهٔ جدید (داده آزمایشی)، آن را با کلیه نمونه‌های موجود در مجموعه داده آموزشی مقایسه و K- نزدیک‌ترین نمونه به آن را استخراج کرده و براساس برتری دسته یا برچسب مربوط به آنها، در مورد دستهٔ داده آزمایشی مزبور تصمیم‌گیری می‌نماید. روش دسته‌بندی KNN در کنار سادگی، دارای دو مشکل اساسی می‌باشد: ۱) حافظه مصرفی و ۲) حجم زیاد محاسبات پردازشی. اولی به دلیل نگه‌داری کل مجموعه داده آموزشی در حافظه و دومی به علت انجام محاسبهٔ فاصله داده جدید با کلیه نمونه‌های آموزشی به وجود می‌آید. هرچه مجموعه داده آموزشی بزرگتر شود، دو مشکل گفته شده بیشتر خود را نشان می‌دهند. ما سعی می‌کنیم در این مقاله با ارائه ایده‌هایی بر دو مشکل فوق فائق آییم و الگوریتم را برای کاربردمان مناسب‌سازی نماییم.

در این پژوهش، ما از الگوریتم K- نزدیک‌ترین همسایه تک کلاسه استفاده می‌کنیم. مهم‌ترین دلیل انتخاب این گونه، کاهش حجم داده آموزشی از طریق نگه‌داری تنها نمونه‌های مثبت (غیر نرمال) می‌باشد. در ادامه، ابتدا مروری به الگوریتم K- نزدیک‌ترین همسایه تک کلاسه نموده و سپس گونه تغییر یافته از آن را پیشنهاد می‌دهیم.

اگر مجموعه داده‌های آموزشی را بـه صـورت $T = \{x_1, x_2, \dots, x_N\}$ در نظـر بگیریم به نحوی که x_i نمونه آموزشی i-اُم در فضای m بعدی باشد، برچسب نمونه آزمایشی x' در دو گام مشخص می‌گردد:

ابتدا فاصله نمونه آزمایشی از تمام نمونـه‌هـای موجـود در مجموعـه آموزشـی محاسبه می‌گردد. برای این منظور از معیار فاصله اقلیدسی که متداول‌تـرین معیـار فاصله است، استفاده می‌گردد:

$$d(x', x_i) = \|x' - x_i\|_{L_2} \quad (۱)$$

منظور از نُرم L_2 یک بردار، مجذور مجموع مربعات مقادیر آن بردار می‌باشد.

K نمونه آموزشی که کمترین فاصله تا داده آزمایشی داشته باشند در مجموعه همسایگی آن قرار می‌گیرند که این مجموعه همسایگی را با NN نمایش می‌دهیم. فرض کنید $near(x)$ نزدیک‌ترین همسایه بـه x در مجموعـه داده آموزشـی باشـد. داده آزمایشی به کلاس مثبت تعلق می‌گیرد اگر:

$$\frac{\sum_{i=1}^{K} d(x', x_i^{NN})}{\sum_{i=1}^{K} d(x_i^{NN}, near(x_i^{NN}))} < \delta \quad (۲)$$

کـه $d(x', x_i^{NN})$ فاصـله داده آزمایشـی از i-اُمـین داده موجـود در مجموعـه همسایگی‌اش و $d(x_i^{NN}, near(x_i^{NN}))$ فاصـله i-اُمـین داده موجـود در مجموعـه همسایگی داده آزمایشی تا نزدیک‌ترین همسایه به خـودش در مجموعـه داده‌هـای آموزشی می‌باشد. نسبت گفته شده باید از آستانهٔ δ کمتر باشـد تـا داده آزمایشـی برچسب "مثبت" دریافت کند (معمولاً $\delta = 1$ در نظر گرفته می‌شود).

در این مقاله، ما به هر داده آموزشی یک فیلد شمارنده نیز اضـافه مـی‌نمـاییم، یعنی $T = \{(x_1, c_1), (x_2, c_2), \dots, (x_N, c_N)\}$. هـدف از افـزودن ایـن فیلـد آن است که هر داده آموزشی بتواند به نماینـدگی از چنـدین نقطـه در مجموعـه داده آموزشی حضور داشته باشد، بنابراین می‌توان بنا بر ظرفیت حافظه، یک محـدودیت برای تعداد داده‌های آموزشی در نظر گرفت و از ایـن طریـق بـه مشـکلات حافظـه مصرفی و حجم محاسبات فائق شد و از طرف دیگر کیفیت داده‌های آموزشی را نیز افزایش داد تا دقت فرآیند دسته‌بندی افزایش یابد. برای این منظور ابتدا رابطه (۲) را به صورت زیر بازنویسی می‌کنیم:

$$\frac{\sum_{i=1}^{K} \frac{1}{d(x', x_i^{NN})}}{\sum_{i=1}^{K} \frac{1}{d(x_i^{NN}, near(x_i^{NN}))}} > \delta' \quad (۳)$$

که در این رابطه، صورت کسر بیانگر میزان شباهت داده آزمایشی به مجموعـه همسایگی‌اش و مخرج کسر میزان نزدیکی نمونه‌های موجود در مجموعه همسایگی داده آزمایشی به سایر داده‌های آموزشی می‌باشد. سپس فیلـد شـمارنده c نیـز بـه عنوان ضریب به آن اضافه می‌کنیم:

$$\frac{\sum_{i=1}^{K} \frac{c_i^{NN}}{d(x', x_i^{NN})}}{\sum_{i=1}^{K} \frac{c_i^{NN(x_i^{NN})}}{d(x_i^{NN}, near(x_i^{NN}))}} > \delta' \quad (۴)$$

که c_i^{NN} شمارنده متناظر با i-اُمین نمونه موجود در مجموعه همسـایگی داده آزمایشی می‌باشد. روش به روز رسانی مقدار c_iها در بخـش بعـدی خواهـد آمـد و همچنین مقدار آستانه δ' در بخش ۴ تعیین می‌شود.

۳-۴- تقویت مجموعه داده آموزشی

همان‌طور که گفته شد، ما یـک مجموعـه اولیـه‌ای از نمونـه‌هـای آموزشـی تولیـد نمودیم اما برای افزایش دقت دسته‌بندی، نیـاز بـه افـزودن نمونـه‌هـای بیشـتر بـه مجموعه داده آموزشی داریم. از طرف دیگر، افزودن نمونـه‌هـای بیشـتر بـه معنـی مصرف حافظه و محاسبات بیشتر در زمان دسته‌بندی می‌باشـد. از همـین رو، یـک فیلد جدید به نام شمارنده (با مقدار اولیه ۱) به هر نمونه آموزشی اضـافه نمـودیم. در زمان دسته‌بندی، اگر داده آزمایشی به عنوان نمونه "مثبت" برچسب زده شـد، به ازای تمام نمونه‌های موجود در مجموعه همسایگی داده آزمایشی، چنانچه فاصله هرکدام از آن نمونه کمتر از آستانه α بود، مقدار فیلد شمارندهٔ آن نمونه آموزشـی به صورت زیر به روز رسانی می‌شود.

$$c_i = c_i + \left(1 - \frac{d(x', x_i^{NN})}{\sum_{j=1}^{K} d(x', x_j^{NN})}\right) \quad (۵)$$

که d بیانگر فاصله داده آزمایشی از نمونه آموزشی می‌باشد. هرچه مقدار فیلـد شمارندهٔ یک داده آموزشی بزرگتر باشد، به این معنا است که این داده به نمایندگی از تعداد بیشتری داده در مجموعه آموزشی حضـور دارد و لـذا از اهمیـت بـالاتری برخوردار است. در نتیجه طبق رابطه (۴) سهم بیشـتری در تعیـین برچسـب یـک دادهٔ جدید خواهد داشت. به همین ترتیب، نمونه‌های آموزشی بـا مقـدار شـمارندهٔ کـوچکتر، نقـش کمتـری را در تعیـین برچسـب دادهٔ جدیـد ایفـا مـی‌کننـد. مقدار α به صورت تجربی برابر با ۰.۲ در نظر گرفته شـده اسـت. نتـایج حاصـل از افزوده شدن این فیلد و بهبود کارایی الگوریتم دسته‌بندی در بخش بعد نشان داده خواهد شد.

سیستم مطرح شده در این مقاله، قابلیت به کارگیری به صـورت بـرخط را دارا می‌باشد. در سیستم برخط، هرگاه موردی به عنوان غیر نرمال تشخیص داده شـود، منجر به نمایش کپچا خواهد شد. در این شرایط، پیشـنهاد مـی‌شـود مـواردی کـه منجر به نمایش کپچا می‌شوند و از طـرف دیگـر کـاربر بـه آنهـا پاسـخ نمـی‌دهـد (ربات‌ها)، به مجموعه آموزشی اضافه گردند. برای این منظـور لازم اسـت ابتـدا بـه دلیل محدودیت حافظه و سربار پردازشی در زمـان دسـته‌بنـدی، آسـتانه‌ای بـرای تعداد داده‌های آموزشی در نظر گرفته شود. سـپس مـی‌تـوان آن داده را تحـت دو شرط زیر به مجموعه آموزشی اضافه نمود:

۱) اگر اندازهٔ مجموعه آموزشی کمتـر از محـدودیت تعیـین شـده باشـد، داده آزمایشی به مجموعه آموزشی افزوده می‌شود.

۲) در غیر این صورت دو نمونه آموزشی از مجموعه نمونـه‌هـای آموزشـی کـه برچسب یکسانی داشته و کمترین فاصله را از یکدیگر دارند یافته و آنهـا را بـا هـم ادغام می‌کنیم تا فضا برای نگه‌داری نمونه آزمایشی جدید باز شود. بـرای ادغـام دو نمونه آموزشی میانگین بردار ویژگی‌های آن دو را محاسبه و فیلدهای شمارندهٔ آنها را با هم جمع کنیم. یعنی دو نمونه (x_i, c_i) و (x_j, c_j) را از مجموعه داده آموزشی حذف و نمونه جدید $\left(\frac{x_i + x_j}{2}, c_i + c_j\right)$ و همچنین نمونه آزمایشی را بـه مجموعـه آموزشی اضافه می‌نماییم.

۴- ارزیابی

گرچه هریک از روش‌های عنوان شده در بخش ۲، با رویکرد متفاوتی بـه مقابلـه بـا کلیک‌های هرز می‌پردازند اما عدم وجود مجموعه دادهٔ عمومی از نشست‌های نرمال و غیر نرمال از رفتار کاربران، موجب دشوار شدن عملیـات ارزیـابی و مقایسـهٔ ایـن روش‌ها می‌شود. لذا عموماً ارزیابی‌ها به بررسی نمونه‌های غیر نرمـال توسـط افـراد

خبره و محاسبهٔ معیارهایی نظیر دقت[۴] محدود می‌شود. ما نیز در این پژوهش، با همین محدودیت مواجه بودیم و تنها امکان محاسبهٔ این معیار را داشتیم، با این حال سعی نمودیم دقت روش‌های پیشنهادی را از جنبه‌های مختلف مورد ارزیابی و مقایسه قرار دهیم.

در ادامه، ابتدا به ارزیابی مجموعه داده آموزشی و مقایسه مجموعه داده اولیه و مجموعه داده آموزشی تقویت شده می‌پردازیم. سپس نحوهٔ انتخاب پارامترها و کارایی الگوریتم دسته‌بند را مورد ارزیابی قرار می‌دهیم و در نهایت دقت روش پیشنهادی را با یکی از آخرین و بهترین کارهای مرتبط مطرح شده مقایسه خواهیم نمود.

۴-۱- اعتبارسنجی متقابل k وجهی[۵]

در ابتدا برای ارزیابی مجموعه داده آموزشی از روش اعتبارسنجی متقابل ۱۰ وجهی استفاده می‌نماییم. در این روش، نمونه‌های آموزشی به صورت تصادفی به ۱۰ بخش مساوی تقسیم می‌شوند، از یک بخش به عنوان داده ارزیابی و از نه بخش دیگر به عنوان داده آموزشی استفاده می‌شود. این فرآیند ۱۰ بار تکرار می‌شود، بنابراین از تمامی نمونه‌ها برای آموزش استفاده می‌شود و هر نمونه نیز یکبار برای ارزیابی مورد استفاده قرار می‌گیرد. در نهایت، مجموع/میانگین نتایج هر تکرار به عنوان تخمین نهایی محاسبه می‌گردد. روش اعتبارسنجی متقابل از معیار دقت دسته‌بندی[۶] بهره می‌گیرد:

$$CA = \frac{TP}{TP + FN} \qquad (۶)$$

که "مثبت صحیح"[۷] بیانگر تعداد نمونه‌هایی است که دستهٔ واقعی آنها "مثبت" بوده و الگوریتم دسته‌بندی نیز دستهٔ آنها را به درستی "مثبت" تشخیص داده است و "منفی کاذب"[۸] معرف تعداد نمونه‌هایی است که دستهٔ واقعی آنها "مثبت" بوده و الگوریتم دسته‌بندی، دستهٔ آنها را به اشتباه "منفی" تشخیص داده است. معیار دقت دسته‌بندی، بیانگر این حقیقت است که دسته‌بند طراحی شده قادر است تا چند درصد از کل مجموعه نمونه‌های آزمایشی را به درستی دسته‌بندی نماید. کمترین مقدار دقت یک دسته‌بند، صفر (ضعیف‌ترین کارایی) و بیشترین مقدار آن، یک (بهترین کارایی) می‌باشد.

مجموعه داده آموزشی اولیهٔ ما شامل ۵۰۰۰ نمونهٔ مثبت که همگی دارای فیلد شمارنده‌ای با مقدار "۱" بودند و الگوریتم فوق را برای مقادیر مختلف K (اندازه مجموعه همسایگی) در الگوریتم KNN تکرار نمودیم. نتایج آن در جدول ۱ نشان داده شده است.

جدول ۱- نتایج اعتبارسنجی متقابل ۱۰ روی مجموعه داده اولیه

	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
مثبت صحیح	۴۶۷۶	۴۷۲۱	۴۷۵۱	۴۷۹۷	۴۸۲۷	۴۸۳۱	۴۸۳۳
منفی کاذب	۳۲۴	۲۷۹	۲۴۹	۲۰۳	۱۷۳	۱۶۹	۱۶۷
دقت (٪)	۹۳/۶۲	۹۴/۴۲	۹۵/۰۲	۹۵/۹۴	۹۶/۵۴	۹۶/۶۲	۹۶/۶۶

در گام بعد، کیفیت داده‌های آموزشی را پس از به روز رسانی فیلد شمارندهٔ آنها مورد بررسی قرار می‌دهیم. برای این منظور، از لاگ یک هفته‌ای تراکنش کاربران (۹۴/۰۹/۱۷ تا ۹۴/۰۹/۲۵) استفاده نموده و آنها را مدل‌سازی نمودیم. سپس این مجموعه را به عنوان داده‌های آزمایشی به سیستم دسته‌بند تک کلاسه تزریق نمودیم. پس از دسته‌بندی داده‌های این مجموعه و به روز رسانی فیلد شمارندهٔ نمونه‌های آموزشی، ما مجدداً مجموعه آموزشی تقویت شده را با کمک روش اعتبارسنجی متقابل ۱۰ وجهی مورد ارزیابی قرار دادیم. نتایج حاصل در جدول ۲ نمایش داده شده است.

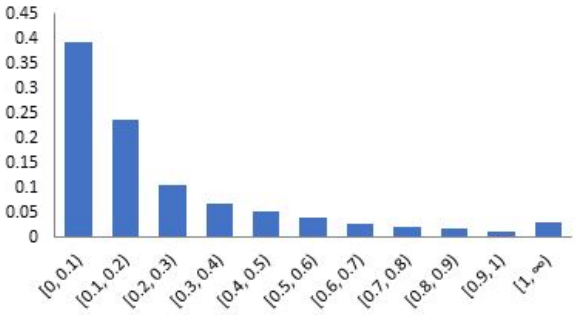
جدول ۲- نتایج اعتبارسنجی متقابل ۱۰ وجهی روی مجموعه داده تقویت شده

	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
مثبت صحیح	۴۸۸۲	۴۹۰۹	۴۹۲۴	۴۹۳۰	۴۹۴۴	۴۹۴۶	۴۹۴۷
منفی کاذب	۱۱۸	۹۱	۷۶	۷۰	۵۶	۵۴	۵۳
دقت (٪)	۹۷/۶۴	۹۸/۱۹	۹۸/۴۸	۹۸/۶	۹۸/۸۸	۹۸/۹۲	۹۸/۹۴

مقایسه نتایج جدول ۱ و ۲، ضمن اینکه بر بهبود دسته‌بند پیشنهادی با افزودن فیلد شمارنده و تقویت داده‌های آموزشی دلالت دارد (افزایش دقت تا ۲٪)، نشان می‌دهد که دقت دسته‌بند پس از تقویت داده‌های آموزشی، به ازای مقادیر مختلف اندازه همسایگی در حدود ۹۸٪ ثابت باقی می‌ماند، لذا ما با توجه به نتایج جدول ۱، از مقدار K=5 برای اندازه مجموعه همسایگی در الگوریتم KNN استفاده نمودیم، بنابراین با این روش، دقت روش پیشنهادی برابر با ۹۶/۵۴٪ خواهد بود.

۴-۲- کارایی الگوریتم دسته‌بندی

در این بخش، کارایی الگوریتم پیشنهادی با استفاده از لاگ یک هفته (۹۴/۰۹/۱۷ تا ۹۴/۰۹/۲۵) که شامل بیش از ۲ میلیون رکورد بود، ارزیابی می‌شود. با توجه به الگوریتم پیشنهادی، به هر نشست پس از دسته‌بندی یک امتیاز تخصیص داده می‌شود (رابطه ۴). این امتیاز در بازهٔ $[0 - \infty)$ متغیر است. ما ابتدا فرکانس امتیاز داده شده به هر نشست را در بازه‌های مختلف به دست می‌آوریم. طول همهٔ بازه‌ها ۰/۱ در نظر گرفته شد ولی جهت محدود کردن تعداد بازه‌ها، امتیازات بزرگتر یا مساوی با ۱ را به صورت مجتمع در یک بازه در نظر می‌گیریم. فرکانس امتیازهای تخصیص داده شده به هر نشست در بازه‌های مختلف در شکل ۲ نشان داده شده است.



شکل ۲- فرکانس امتیازهای حاصل از دسته‌بندی در بازه‌های مختلف

با توجه به الگوریتم پیشنهادی، می‌دانیم هرچه امتیاز محاسبه شده بیشتر باشد، احتمال غیر نرمال بودن آن نشست بیشتر می‌شود. ما برای آزمایشات بیشتر، از بین نشست‌هایی که امتیاز بالاتر از ۰/۵ داشتند، به صورت تصادفی ۲۰۰ نشست

را انتخاب و دقت الگوریتم پیشنهادی را مورد بررسی قرار دادیم که نتایج آن در جدول ۳ نمایش داده شده است.

جدول ۳- دقت الگوریتم پیشنهادی در بازه‌های امتیازی مختلف

محدوده امتیاز	تعداد نشست‌ها	تعداد غیر نرمال‌ها	دقت (٪)
[۰.۵-۰.۶)	۴۱	۲	۴/۸۷
[۰.۶-۰.۷)	۲۳	۳	۱۳/۰۴
[۰.۷-۰.۸)	۱۸	۵	۲۷/۷۷
[۰.۸-۰.۹)	۱۹	۱۷	۸۹/۴۷
[۰.۹-۱)	۱۵	۱۴	۹۳/۳۳
[۱-∞)	۸۴	۸۳	۹۸/۸۰

با توجه به اینکه شناسائی موارد هرز نیاز به دقت بالایی دارد، لذا ما نشست‌هایی که امتیاز بالاتر از ۰/۸ داشته‌اند ($\delta' = 0.8$) را به عنوان نشست غیر نرمال در نظر گرفتیم. با این حساب، دقت روش برابر با ۹۶/۶۱٪ می‌شود که از نسبت ۱۱۴ نشست غیر نرمال درست تشخیص داده شده به ۱۱۸ نشست مورد بررسی در بازهٔ [۰/۸-∞] به دست می‌آید. در نتیجه کلیه کلیک‌های انجام شده در این نشست‌ها به عنوان کلیک هرز قلمداد می‌شوند.

۴-۳- مقایسه با کارهای قبلی

در نهایت، کارایی الگوریتم پیشنهادی با دو روش اخیر مطرح شده جهت شناسائی کلیک‌های هرز [۲۰،۲۱] نیز مقایسه گردید.

در روش معرفی شده در [۲۰]، ابتدا نشست‌های کاربران به صورت توالی‌های دوتایی از (نوع فعالیت کاربر و شماره صفحه) و با کمک زنجیرهٔ مارکوف مدل می‌شود. سپس فاصله هر نشست از میانگین نشست‌ها با معیار فاصله ماهالانوبیس محاسبه می‌گردد. آنگاه نشست‌هایی که در یک درصد بالایی توزیع فاصله نشست‌ها قرار دارند، به عنوان نشست غیر متعارف در نظر گرفته می‌شوند. همچنین، در الگورتیم دیگری که در سال ۲۰۱۴ توسط لی و همکارانش [۲۱] مطرح گردید، ابتدا نشست‌های کاربران به صورت توالی‌های سه‌تایی از (نوع فعالیت، هدف فعالیت و اختلاف زمانی آن با فعالیت قبلی) مدل گردیده و سپس الگوریتم گراف دو بخشی کاربر- نشست برای شناسائی تعداد بیشتری از نشست‌های مشکوک اعمال می‌شود. در انتها نشست‌هایی که امتیازی بالاتر از ۰/۹ داشته‌اند به عنوان نشست غیر نرمال در نظر گرفته می‌شوند. از بین نتایج حاصل از این دو روش به همراه روش پیشنهادی، به صورت تصادفی ۲۰۰ نمونه از آنها را انتخاب و درستی تشخیص آنها را بررسی نمودیم. نتایج و دقت هر سه روش در جدول ۴ نشان داده شده است.

جدول ۴- مقایسه با روش‌های قبلی

	دقت (٪)
روش نشست‌های متعارف/غیر متعارف [۲۰]	۸۹
روش گراف دوبخشی[۲۱]	۹۱.۵
روش پیشنهادی	۹۶/۶۱

مقایسهٔ دقت محاسبه شده به ازای روش‌های قبلی و روش پیشنهادی، نشان می دهد که روش پیشنهاد شده در این مقاله در حدود ۵/۱ درصد عملکرد بهتری روی داده‌های ما داشته است. البته باید به این نکته نیز توجه نمود که روش‌های قبلی بیشتر به صورت برون‌خط عمل می‌کنند، در حالیکه روش پیشنهادی در این مقاله قادر است تا به صورت لحظه‌ای و با مصرف حافظهٔ پائین‌تر و انجام محاسباتی ساده‌تر به شناسائی کلیک‌های هرز بپردازد.

۵- نتیجه‌گیری

در این مقاله، یک روش جدید و کارآمد مبتنی بر دسته‌بندی جهت شناسائی نشست‌ها و کلیک‌های هرز ارائه نمودیم. ما در ابتدا ویژگی‌های مهمی را که به تمایز رفتار کاربران نرمال و ربات‌ها کمک می‌کرد معرفی کردیم و سپس الگوریتم دسته‌بندی جدیدی که مبتنی بر الگوریتم KNN می‌باشد را پیشنهاد دادیم که در آن با افزودن یک پارامتر سادهٔ "شمارنده" توانستیم بر مشکلات الگوریتم اولیه KNN که شامل حافظه مصرفی و حجم زیاد محاسبات بود، فائق آییم. اما از طرف دیگر با ایجاد مکانیزم به‌روزرسانی مجموعه آموزشی، سعی نمودیم همواره امکان افزودن نمونه‌های جدید به سیستم وجود داشته باشد. در بخش ارزیابی نشان دادیم که الگوریتم پیشنهادی می تواند کلیک‌های هرز را با دقتی بیش از ۹۶٪ تشخیص دهد که در مقایسه با روش‌های قبلی بهبود مناسبی از خود نشان می‌دهد. ما قصد داریم در آینده‌ای نزدیک، الگوریتم مطرح شده را در محیط برخط آزمایش و نتایج آن را در کارهای بعدی گزارش نماییم.

تشکر و قدردانی

از مجموعهٔ موتور جستجوی پارسی‌جو جهت تأمین داده‌های مورد نیاز در این پژوهش تشکر می‌نماییم.

مراجع

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford InfoLab, 1999.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, "Link-Based Characterization and Detection of Web Spam," *Proc. Intl Workshop on Adversarial Information Retrieval on the Web*, 2006.

[3] Y. Liu, R. Cen, M. Zhang, S. Ma, and L. Ru, "Identifying Web Spam with User Behavior Analysis," *Proc. Intl Workshop on Adversarial Information Retrieval on the Web*, pp. 9–16, 2008.

[4] A. Karasaridis, B. Rexroad, and D. Hoeflin, "Wide-scale Botnet Detection and Characterization," *Proc. Conf. First Workshop on Hot Topics in Understanding Botnets*, pp. 7, 2007.

[5] Z. Dou, R. Song, X. Yuan, and J.-R. Wen, "Are Click-through Data Adequate for Learning Web Search Rankings?," *Proc. ACM Conf. Information and Knowledge Management*, pp. 73–82, 2008.

[6] I. A. Board, "Internet Advertising Revenue Report," https://www.iab.com, 2015.

[7] D. Szetela, and J. Kerschbaum, *Pay-Per-Click Search Engine Marketing: An Hour a Day*, USA: SYBEX Inc.,

2010.

[8] N. Daswani, and M. Stoppelman, "The Anatomy of Clickbot.A," *Proc. Conf. First Workshop on Hot Topics in Understanding Botnets*, pp. 11, 2007.

[9] B. Miller, P. Pearce, C. Grier, C. Kreibich, and V. Paxson, "What's Clicking What? Techniques and Innovations of Today's Clickbots," *Proc. Intl Conf. Detection of Intrusions and Malware, and Vulnerability Assessment*, pp. 164–183, 2011.

[10] S. A. Alrwais, A. Gerber, C. W. Dunn, O. Spatscheck, M. Gupta, and E. Osterweil, "Dissecting Ghost Clicks: Ad Fraud via Misdirected Human Clicks," *Proc. Conf. Computer Security Applications*, pp. 21–30, 2012.

[11] P. Pearce, and et. al., "Characterizing Large-Scale Click Fraud in ZeroAccess," *Proc. ACM SIGSAC Conf. Computer and Communications Security*, pp. 141–152, 2014.

[12] B. Stone-Gross, R. Stevens, A. Zarras, R. Kemmerer, C. Kruegel, and G. Vigna, "Understanding Fraudulent Activities in Online Ad Exchanges," *Proc. ACM SIGCOMM Conf. Internet Measurement*, pp. 279–294, 2011.

[13] V. Dave, S. Guha, and Y. Zhang, "ViceROI: Catching Click-spam in Search Ad Networks," *Proc. ACM SIGSAC Conf. Computer andCommunications Security*, pp. 765–776, 2013.

[14] R. Oentaryo, and et. al., "Detecting Click Fraud in Online Advertising: A Data Mining Approach," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 99–140, Jan. 2014.

[15] B. Kitts, J. Y. Zhang, G. Wu, W. Brandi, J. Beasley, K. Morrill, J. Ettedgui, S. Siddhartha, H. Yuan, F. Gao, P. Azo, and R. Mahato, "Click Fraud Detection: Adversarial Pattern Recognition over 5 Years at Microsoft," *Springer Intl Pub. Real World Data Mini. Apps.*, vol. 17, pp. 181–201, 2015.

[16] H. Haddadi, "Fighting Online Click-fraud Using Bluff Ads," *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 2, pp. 21–25, Apr. 2010.

[17] B. Kitts, J. Y. Zhang, A. Roux, and R. Mills, "Click Fraud Detection with Bot Signatures," *Proc. IEEE Intl Conf. Intelligence and Security Informatics*, pp. 146–150, 2013.

[18] F. Yu, Y. Xie, and Q. Ke, "SBotMiner: Large Scale Search Bot Detection," *Proc. ACM Intl Conf. Web Search and Data Mining*, pp. 421–430, 2010.

[19] H. Kang, K. Wang, D. Soukal, F. Behr, and Z. Zheng, "Large-scale Bot Detection for Search Engines," *Proc. Intl Conf.World Wide Web*, pp. 501–510, 2010.

[20] N. Sadagopan, and J. Li, "Characterizing Typical and Atypical User Sessions in Clickstreams," *Proc. Intl Conf. World Wide Web*, pp. 885–894, 2008.

[21] X. Li, M. Zhang, Y. Liu, S. Ma, Y. Jin, and L. Ru, "Search Engine Click Spam Detection Based on Bipartite Graph Propagation," *Proc. ACM Intl Conf. Web Search and Data Mining*, pp. 93–102, 2014.

[22] G. Buehrer, J. W. Stokes, and K. Chellapilla, "A Large-scale Study of Automated Web Search Traffic," *Proc. Intl Workshop on Adversarial Information Retrieval on the Web*, pp. 1–8, 2008.

[23] R. A. Costa, R. J. G. B. de Queiroz, and E. R. Cavalcanti, "A Proposal to Prevent Click-Fraud Using Clickable CAPTCHAs," *Proc. IEEE Intl Conf. Software Security and Reliability Companion*, pp. 62–67, 2012.

[24] T. Cover, and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Inf. Theor.*, vol. 13, no. 1, pp. 21–27, Sep. 2006.

مهدیه فلاح مدرک کارشناسی خود را در رشتهٔ مهندسی فناوری اطلاعات در دانشگاه صنعتی اصفهان اخذ نمود. پس از آن در مقطع کارشناسی‌ارشد در همان رشته با گرایش شبکه‌های کامپیوتری به ادامهٔ تحصیل پرداخت. وی به موتورهای جستجو و شبکه‌های کامپیوتری علاقهٔ زیادی دارد و هم‌اکنون در پارسی‌جو، اولین موتور جستجوی ایرانی مشغول به فعالیت می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

m.fallah@stu.yazd.ac.ir

سجاد ظریف‌زاده دکتری خود را در رشته مهندسی کامپیوتر از دانشگاه تهران در سال ۱۳۹۱ دریافت نمود. وی هم‌اکنون استادیار و عضو هیئت علمی گروه مهندسی کامپیوتر دانشگاه یزد می‌باشد. زمینه‌های تحقیقاتی اصلی وی در مورد سرویس‌های مبتنی بر وب و همچنین سیستم‌ها و شبکه‌های کامپیوتری می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

szarifzadeh@yazd.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۰۵/۱۲

تاریخ اصلاح: ۱۳۹۵/۰۶/۰۴

تاریخ قبول شدن: ۱۳۹۵/۰۶/۱۴

نویسنده مرتبط: مهدیه فلاح، دانشکده برق و کامپیوتر، دانشگاه یزد، یزد، ایران.

[1] Domain Name Server
[2] Ad Exchanger
[3] K-Nearest Neighbor (KNN)
[4] Precision
[5] K-Fold Cross-Validation
[6] Classification Accuracy
[7] True Positive (TP)
[8] False Negative (FN)