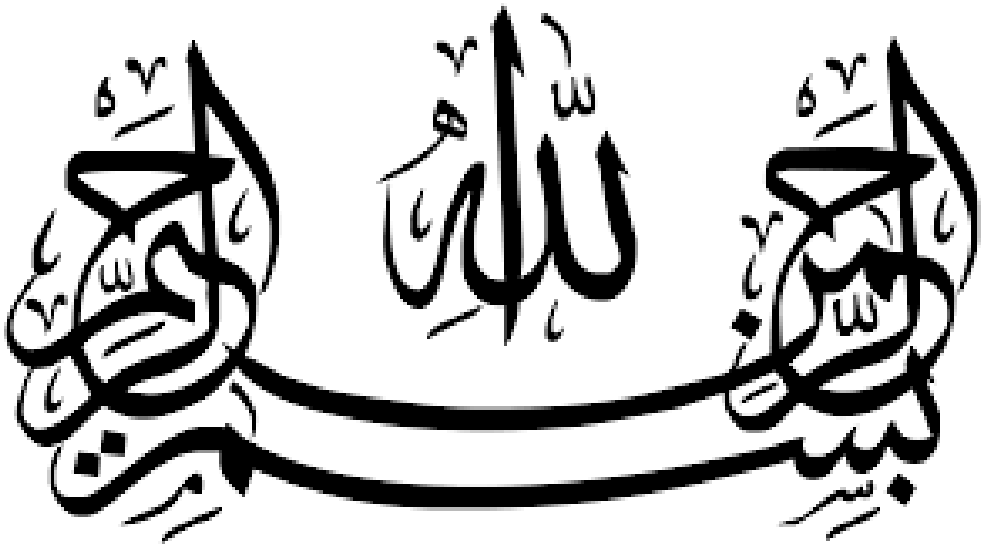


دانشگاه یزد

دانشکده مهندسی برق و کامپیوتر

گروه مهندسی کامپیوتر

پایان نامه

برای دریافت درجه کارشناسی ارشد

در رشته مهندسی فناوری اطلاعات - گرایش شبکه‌های کامپیوتری

عنوان

شناسائی کلیک‌های هرز در فضای وب فارسی

استاد راهنما

دکتر سجاد ظریف‌زاده

استاد مشاور

دکتر ولی درهمی

پژوهش و نگارش

مهدیه فلاح

اسفند ۱۳۹۴

نام بعضی نفرات رزق روحم شده، جرأتم می‌بخشد و روشنم می‌دارد...

تقدیم به آنان

تقدیم به خداوند،

که بزرگترین امید و یاور در لحظه لحظه زندگیم است

تقدیم به مادرم،

دریای بی‌کران فداکاری و عشق که وجودم برایش همه رنج بود و وجودش برایم همه مهر

تقدیم به پدرم،

که عالمانه به من آموخت تا چگونه در عرصه زندگی، ایستادگی را تجربه نمایم

و تقدیم به همه کسانی

که دوستشان دارم...

سپاس خدای را،

که سخنوران در ستودن او بمانند و شمارندگان، شمردن نعمت‌های او ندانند و کوشندگان، حق او را گزاردن نتوانند. سپاس او را در برابر عطا و احسانش....

بر خود واجب می‌دانم از جناب آقای دکتر سجاد ظریف‌زاده که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ مساعدتی در این عرصه بر من دریغ ننموده و زحمت راهنمایی این پژوهش را بر عهده گرفتند و همچنین از استاد صبور و فرهیخته، جناب آقای دکتر ولی دهی که زحمت مشاوره این پژوهش را متقبل شدند؛ کمال تشکر و قدردانی را داشته باشم. باشد که این خرد ترین، بخشی از زحمات آنان را سپاس گوید.

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 همین رو، برخی وب‌سایت‌ها 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 است، اصطلاحاً به آنها کلیک‌های هرز گفته می‌شود. برای این منظور، وب‌سایت‌ها معمولاً از برنامه‌های نرم‌افزاری به نام "ربات‌ها" استفاده می‌کنند تا به صورت خودکار و توزیع شده به ارسال تعداد زیادی پرس‌وجو و همچنین انجام کلیک‌های هرز بپردازند. در این پژوهش، روش‌های 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. این ویژگی‌ها به نحوی انتخاب شده‌اند که بتوانند جنبه‌های مختلفی از رفتارهای نرمال و غیر نرمال را پوشش داده و تا حد امکان آنها را از یکدیگر تمایز دهند. سپس، در گام بعد با اعمال الگوریتم‌های دسته‌بندی پیشنهادی که شامل یک الگوریتم دسته‌بندی دو کلاسه و یک الگوریتم تک کلاسه می‌باشد، اقدام به شناسائی نشست‌های غیر نرمال و در نتیجه کلیک‌های هرز می‌نماییم. روش‌های مطرح شده با لاگ واقعی یک موتور جستجو فارسی مورد تحلیل قرار گرفته است. نتایج بررسی‌ها نشان می‌دهد که روش‌های پیشنهادی می‌توانند کلیک‌های هرز را با دقتی بیش از ۹۶ درصد تشخیص دهند که در مقایسه با کارهای قبلی تا ۳/۵ درصد بهبود از خود نشان می‌دهد.

کلمات کلیدی: کلیک هرز، شناسائی ربات‌ها، یادگیری ماشین، K-نزدیک‌ترین همسایه

فهرست مطالب

عنوان صفحه

فهرست جداول

عنوان صفحه

فهرست اشکال

عنوان صفحه

فصل اول: مقدمه

۱-۱ تعریف مسئله

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 نموده و به کاربر نمایش می‌دهند.

در دهۀ اخیر، موتورهای جستجو به منظور بهبود کیفیت نتایج بازگشتی 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. این امر می‌تواند موجب سوء استفاده از موتورهای جستجو و دست‌کاری صفحۀ نتایج به منظور بالا بردن رتبۀ برخی صفحات خاص و یا احیاناً خرابکاری شود. 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. بنابراین، 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 تأثیر منفی روی تصمیم گیری‌هایی که بر اساس سابقه و بازخورد کاربران گرفته می‌شود، به موتور جستجو صدمه بزند.

از سوی دیگر، درآمد اصلی سرویس‌های رایگانی نظیر موتورهای جستجو از سیستم تبلیغات آنلاین آنها می‌باشد. افزایش روزافزون کاربران اینترنتی نیز به رونق این کسب و کار کمک شایانی نموده به نحوی که درآمد حاصل از تبلیغات آنلاین در سال ۲۰۱۴ به ۴۹/۵ میلیارد دلار رسیده است که این مقدار نسبت به سال قبل خود، بیشتر از ۱۵٪ رشد داشته است [۳].

سیستم تبلیغات آنلاین از سه مؤلفه اصلی تشکیل می‌شود: ۱) صاحبان تبلیغ (تبلیغ‌کنندگان) که جهت معرفی محصولات و یا وب‌سایت خود، اقدام به ثبت نام و ایجاد حساب

کاربری در شبکه‌های تبلیغاتی می‌نمایند. آنها پس از درج تبلیغ خود، بر اساس بودجه‌ای که برای تبلیغات در نظر گرفته‌اند، حساب خویش را شارژ می‌کنند. ۲) منتشرکنندگان که وظیفه میزبانی و نمایش تبلیغات را انجام می‌دهند. موتورهای جستجو می‌توانند خود به عنوان منتشرکننده به نمایش تبلیغات در کنار نتایج جستجو به کاربران بپردازند و یا اینکه از وب‌سایت‌ها و یا برنامه‌های کاربردی که با ثبت نام در شبکه تبلیغ اعلام آمادگی نموده‌اند، برای نمایش تبلیغات گرافیکی استفاده نمایند. ۳) شبکه‌های تبلیغ که مدیریت ارتباطات بین تبلیغ‌کنندگان و منتشرکنندگان و همچنین مسئولیت انتخاب تبلیغ مناسب جهت نمایش به کاربران را به عهده دارند.

شبکه‌های تبلیغاتی معمولاً بر مبنای مدل "پرداخت به ازای هر کلیک"[1] [۴] فعالیت می‌نمایند یعنی هر زمان روی یک تبلیغ کلیک شود، مبلغی از شارژ تبلیغ‌کننده کسر می‌گردد. این مبلغ در میان تبلیغات مختلف، متفاوت است و به نوع و محتوای تبلیغ بستگی دارد. بر اساس مدل گفته شده، هرچه تعداد کلیک‌های بیشتری روی یک تبلیغ صورت بگیرد، بودجه مربوط به آن تبلیغ زودتر به اتمام می‌رسد. مسئله "کلیک‌های هرز"[2] در اینجا ظهور پیدا می‌کند که می‌توان به سادگی با انجام کلیک روی یک تبلیغ خاص، بودجهٔ آن را تمام نمود، بدون اینکه واقعاً کاربری وجود داشته باشد. لذا، چون این کلیک‌ها توسط کاربران واقعی و علاقمند به تبلیغ صورت نمی‌گیرد، اصطلاحاً به آنها کلیک‌های هرز گفته می‌شود.

کلیک‌های هرز در سیستم تبلیغات آنلاین به دو نوع تقسیم می‌شوند: تورم کلیک و کلیک‌های رقابتی. در حالت اول، منشرکنندگان با انگیزهٔ کسب درآمد بیشتر به انجام هر چه بیشتر کلیک روی تبلیغات نمایش داده شده در وب‌سایت خویش می‌پردازند و در حالت دوم، رقبای یک تبلیغ‌کننده با انگیزهٔ تمام کردن بودجه تبلیغات رقیب خود به این کار مبادرت می‌ورزند. ارتباط مستقیم این نوع از کلیک‌های هرز با مسائل مالی، موجب شده است که در مجامع علمی اصطلاحاً به آنها "کلیک کلاهبردارانه"[3] نیز گفته شود اما ما در این پژوهش، همچنان از آنها با عنوان کلیک هرز یاد می‌کنیم. گرچه با وجود این کلیک‌ها، موتورهای جستجو باز هم به ازای هر کلیک درآمد

[1] Pay Per Click (PPC)
[2] Click Spam
[3] Click Fraud

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.

گرچه مسئله کلیک‌های هرز در سایر سیستم‌های آنلاین نظیر سیستم‌های توصیه‌گر و سایت‌های خبری نیز 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 بیشتری برخوردار است. لذا، ما در این پژوهش صرفاً بر روی مسئله شناسائی کلیک‌های هرز در موتورهای جستجو تمرکز می‌کنیم.

کلیک‌های هرز در موتورهای جستجو بر اساس هدفشان به دو نوع تقسیم می‌شوند:

- کلیک‌های هرز روی نتایج اصلی موتورهای جستجو با هدف افزایش رتبهٔ یک وب‌سایت در صفحهٔ نتایج
- 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 غیرنرمال 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 [۵-۷]. بنابراین، شناسائی آنها بسیار دشوار و پیچیده شده است. به عنوان مثال، در مقاله [۷] 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

[1] Domain Name Server

ترافیک‌های غیرنرمال و تفکیک آنها از ترافیک کاربران واقعی روی آورده‌اند.

در این پژوهش، ما سعی می‌کنیم جنبه‌های مختلفی از رفتار ترافیک‌های غیرنرمال را به صورت مجموعه‌ای از ویژگی‌ها در سه سطح نشست، کاربر و آدرس IP با یکدیگر ترکیب نموده و با کمک یک تکنیک دسته‌بندی جدید که گونه‌ای تغییر یافته از الگوریتم K-نزدیک‌ترین همسایه می‌باشد، به شناسائی کلیک‌های هرز بپردازیم. همچنین اکثر روش‌هایی که تاکنون جهت شناسائی کلیک‌های هرز در موتورهای جستجو پیشنهاد شده‌اند، به صورت برون‌خط[1] کار می‌کنند اما سیستم پیشنهاد شده در این تحقیق، قادر است به صورت برخط[2] به شناسائی کلیک‌های هرز بپردازد.

۱-۲ ساختار پژوهش

ساختار این پایان‌نامه در ادامه به شرح زیر می‌باشد. در فصل دوم، مروری کلی بر روی مقالات ارائه شده در زمینهٔ شناسائی کلیک‌های هرز در دو حوزهٔ شبکه‌های تبلیغاتی و موتورهای جستجو می‌نماییم. مسئله تولید داده‌های آموزشی برچسب‌دار در فصل سوم مورد بررسی قرار می‌گیرد و سپس روش پیشنهاد شده جهت شناسائی کلیک‌های هرز را مطرح می‌کنیم. بیان ارزیابی روش پیشنهادی و مقایسهٔ آن با روش‌های قبلی در فصل چهارم انجام می‌گیرد. در نهایت، فصل هفتم را به نتیجه‌گیری و بیان کارهای بعدی در این زمینه اختصاص داده‌ایم.

[1] Offline
[2] Online

فصل دوم: سابقه تحقیق

۲-۱ مقدمه

همانطور که در فصل قبل گفته شد، کلیک‌های هرز به دو دسته تقسیم 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 با هدف کسب درآمد بیشتر توسط منتشرکنندگان و یا تمام کردن بودجهٔ یک تبلیغ‌کنندهٔ خاص. موتورهای جستجو باید بتوانند هر دو نوع را تشخیص داده و از ترافیک کاربران نرمال تمایز دهند. از اینرو، محققان زیادی بر روی این مسئله تمرکز کرده و به ارائه راهکارهایی جهت شناسائی کلیک‌های هرز پرداخته‌اند. پژوهش‌های انجام شده در این زمینه نیز خود به دو بخش تقسیم می‌شوند: ۱) شناسائی کلیک‌های هرز در شبکه‌های تبلیغاتی و ۲) شناسائی کلیک‌های هرز در موتورهای جستجو. در دو بخش آتی، به ترتیب به مرور تعدادی از مقالات ارائه شده در هریک از زمینه‌ها می‌پردازیم.

البته باید ذکر گردد که روش‌های شناسائی کلیک‌های هرز در شبکه‌های تبلیغاتی مستقیماً به پژوهش انجام گرفته در این پایان‌نامه مربوط نمی‌شوند اما ما برای تکمیل بحث، به معرفی آنها نیز پرداخته‌ایم.

۲-۲ شناسائی کلیک‌های هرز در شبکه‌های تبلیغات

در دههٔ اخیر با افزایش روزافزون کاربران اینترنتی و داغ شدن بحث تبلیغات آنلاین، شناسائی کلیک‌های هرز روی تبلیغات به عنوان یک چالش اساسی در بسیاری از مجامع علمی مورد بررسی و تحقیق قرار گرفته است. همانطور که پیش‌تر، نیز گفته شد کلیک‌های هرز در شبکه‌های تبلیغاتی می‌تواند توسط منتشرکنندگان با انگیزهٔ کسب درآمد بیشتر و یا رقبای یک تبلیغ‌کننده با هدف تمام کردن بودجهٔ رقیب خود صورت گیرد. پژوهش‌های انجام شده در حوزهٔ شناسائی کلیک‌های هرز در شبکه‌های تبلیغاتی عمدتاً به شناسائی منتشرکنندگان کلاهبردار معطوف شده است. به این ترتیب، با شناسائی منتشرکنندگان متقلب تمام کلیک‌های انجام گرفته از سوی آنها به عنوان کلیک هرز تلقی می‌گردد.

هریک از روش‌های مطرح شده در این زمینه، از منظر خاصی به مسئله نگاه کرده و

تکنیک‌هایی جهت تشخیص پیشنهاد داده‌اند. روش‌های تشخیص ناهنجاری، خوشه‌بندی منتشرکنندگان و روش‌های مبتنی بر درآمد حاصل شده برای هر منتشرکننده از جمله مهم‌ترین روش‌هایی هستند که در این حوزه مطرح شده‌اند و ما در اینجا به معرفی آنها می‌پردازیم.

۲-۲-۱ ناهنجاری در میان منتشرکنندگان

در مقاله [۸]، پژوهشگران از روشی مشابه سیستم‌های تشخیص نفوذ سنتی استفاده نمودند. برای این منظور، به ازای هر یک از درخواست‌های منجر به نمایش یا کلیک روی تبلیغات که به سمت شبکه تبلیغ می‌آید، ویژگی‌های زیر استخراج شده و برای مدل کردن مشخصات ترافیکی آنها استفاده می‌شود:

- تعداد دفعات نمایش تبلیغ به ازای هر کوکی (ذخیره شده در مرورگر کاربر)
- نرخ کلیک از هر کوکی
- درآمد منتشرکننده به ازای هر کوکی
- تعداد آدرس‌های IP یکتا به ازای هر کوکی
- تعداد دفعات نمایش تبلیغ به ازای هر آدرس IP
- نرخ کلیک از هر کوکی
- درآمد منتشرکننده به ازای هر آدرس IP

برای هریک از ویژگی‌های عنوان شده، یک حد آستانه تعریف می‌شود که مقدار آن به صورت پویا بر اساس میانگین و واریانس هر ویژگی در بازه زماننی یک ساعته محاسبه و روی داده‌های یک ساعت بعد اعمال می‌شود. در هر بازه زمانی (یک ساعت)، کوکی‌ها و آدرس‌های IP که آستانه‌های تعریف شده تجاوز می‌نمایند، شناسائی شده و تمام درخواست‌هایی که از سوی این آدرس‌ها یا کوکی‌ها می‌آیند با عنوان "درخواست‌های مشکوک" استخراج می‌شوند. در گام بعد، منتشرکنندگانی که این درخواست‌ها از سوی آنها آمده است، مشخص می‌گردند. به این ترتیب می‌توان به ازای هر منتشرکننده، نسبت درخواست‌های مشکوک به کل درخواست‌های آمده از سوی وی را (هم در بازه زمانی یک ساعته و هم در کل بازه زمانی) محاسبه نمود. مقادیر حاصل به

عنوان معیاری از رفتار ناهنجار برای هر منتشرکننده در نظر گرفته می‌شود. هرچه این مقدار بیشتر باشد، احتمال متقلب بودن منتشرکننده نیز بیشتر است.

۲-۲-۲ خوشه‌بندی منتشرکنندگان

در سال ۲۰۱۰، حدادی ایدهٔ به کارگیری تبلیغات بلوف را جهت تشخیص کلیک‌های هرز و به طور خاص کلیک‌هایی که به صورت خودکار توسط ربات‌ها انجام می‌گیرند را مطرح نمود [۹]. تفاوت اصلی تبلغات بلوف با تبلیغات معمولی در متن آنها است. متن تبلیغات معمولی، مجموعه‌ای از کلمات مرتبط با یکدیگر است که به شکل‌گیری یک عبارت معنی‌دار منجر می‌شود، در حالی که متن تبلیغات بلوف، مجموعه‌ای تصادفی از کلمات می‌باشد که از فرهنگ لغت انتخاب شده‌اند. بنابراین، یک کاربر با مشاهدهٔ این نوع تبلیغات، تنها ترکیبی از کلمات مختلف را می‌بیند بدون اینکه معنی مشخصی از آنها برداشت نماید. در مقاله نشان داده شده است که افراد معمولی (کاربران معتبر) تمایلی به کلیک روی این تبلیغات ندارند و عمده کلیک‌ها از سوی ربات‌ها انجام می‌گیرد که هنوز به این درجه از هوشمندی جهت تشخیص نرسیده‌اند. شکل ۲-۱ نمونه‌ای از تبلیغات بلوف در موتورهای جستجو را نشان می‌دهد.

شکل ۲-۱: نمونه‌ای از تبلیغات بلوف در موتورهای جستجو [۱۰]

دو سال بعد، داو و همکارانش با بهره‌گیری از این ایده، روشی پیشنهاد دادند که هر یک از تبلیغ‌کنندگان، خو بتوانند نرخ کلیک‌های هرز روی تبلیغاتشان را اندازه‌گیری نمایند [۱۰]. در این روش، محققان تعداد تبلیغ معمولی و چند تبلیغ بلوف ایجاد کرده و به عنوان تبلیغ‌کننده در بزرگ‌ترین شبکه‌های تبلیغات نظیر گوگل و یاهو ثبت نام نمودند. علاوه بر این، برای هر یک از تبلیغات، صفحه‌ای آماده نمودند که کاربران با کلیک روی تبلیغات به آنها ارجاع داده شوند.

همچنین، آنها چندین صفحه میانی طراحی کردند تا قبل از نمایش صفحهٔ اصلی تبلیغ به کاربر نشان داده شود. بنابراین، زمانی‌که کاربر روی یک تبلیغ کلیک می‌کند یا مستقیماً وب‌سایت تبلیغ‌کننده برگزاری می‌شود یا ابتدا یک صفحهٔ میانی به کاربر نشان داده می‌شود و بعد وارد وب‌سایت تبلیغ‌کننده می‌شود. در صفحه میانی ممکن است یکی از سه چالش زیر پیش روی کاربر قرار می‌گیرد:

- با نشان دادن پیغام "صفحه در حال بارگزاری است..."، کاربر می‌بایست به مدت ۵ ثانیه منتظر بماند.
- از کاربر خواسته می‌شود روی یک لینک کلیک کند تا وارد وب‌سایت اصلی شود.
- از کاربر خواسته شود ابتدا یک کد کپچا[1] وارد نماید و سپس وب‌سایت اصلی نشان داده می‌شود.

فرض شده است که صفحه میانی حجم زیاد از کلیک‌های هرز را متوقف می‌سازد، ولی همچنان درصدی از ترافیک‌های هرز و همچنین ترافیک کاربران معتبر به کار خود ادامه می‌دهند. بررسی‌های انجام گرفته نشان می‌دهد که وجود صفحهٔ میانی تأثیری روی ترافیک کاربران معتبر نداد و آنها همچنان با وجود این چالش‌ها، ادامه داده و به وب‌سایت تبلیغ‌کننده خواهند آمد. بنابراین، هر یک از تبلیغ‌کننده‌ها می‌توانند از چنین مکانیزمی استفاده نمایند و با مقایسه تعداد کلیک‌های رسیده از شبکه‌های مختلف تبلیغات به صفحهٔ نهایی و تعداد کلیک‌هایی که هر شبکه تبلیغ از اعتبار آنها کم کرده است، نرخ کلیک‌های هرز روی تبلیغات خود را به دست آوردند.

از طرف دیگر، گفته شد که معمولاً کاربران معتبر روی تبلیغات بلوف کلیک نمی‌کنند، لذا کلیک‌های انجام شده روی این تبلیغات به صورت دقیق‌تر مورد بررسی قرار گرفتند. برای این منظور، از یک گراف خوشه‌بندی استفاده نمودند زیرا حجم کلیک‌های انجام شده روی تبلیغات بلوف نسبتاً زیاد می‌باشد. برای ایجاد گراف خوشه‌بندی، به ازای هر منتشرکننده، برداری از ویژگی‌های سطح شبکه‌ای (نظیر آدرس IP، ثبت‌کننده دامنه و اطلاعات Whois) و ویژگی‌های سطح HTTP (مانند فیلد ارجاع‌دهنده از هر درخواست) ایجاد و به هر ویژگی یک وزن تخصیص

[1] Captcha

داده می‌شود. سپس، به ازای هر جفت منتشرکننده، از معیار شباهت کسینوس بین دو بردار جهت شباهت‌سنجی منتشرکنندگان استفاده می‌گردد؛ در صورتی که مقدار حاصل از یک آستانه بیشتر باشد، بین دو منتشرکننده یک یال وصل می‌شود. در نهایت، خوشه‌های چگال به دست آمده از عملیات خوشه‌بندی، هر یک معرف حملات مختلفی می‌باشند که از همکاری چندین منتشرکننده با یکدگیر حاصل شده‌اند.

۲-۲-۳ درآمدزایی منتشرکنندگان

دسته دیگری از روش‌های شناسائی کلیک‌های هرز از درآمد تولید شده برای هر منتشرکننده جهت شناسائی منتشرکنندگان متقلب استفاده می‌نمایند. اساس طراحی این دسته از روش‌ها بر این مبنا است که میزان بازگشت سرمایه منتشرکنندگان متقلب نسبت به منتشرکنندگان درستگار بیشتر است. در روش مطح شده در [۱۱]، محققان ابتدا به بررسی سیستم هزینه-درآمد در شبکه‌های رباتی می‌پردازند. بر این اساس، چهار متغیر که در میزان سود یک حمله‌کننده نقش دارند، مطرح می‌شوند: ۱) هزینه ثابت صرف شده مانند هزینه اجارهٔ یک شبکه رباتی، ۲) هزینه افزایشی که به ازای هر کلیک افزوده می‌شود، نظیر هزینه کلیک‌های انجام شده در شبکه تبلیغ اول در حملات معامله به سود (در این نوع حملات، خود منتشرکنندگان نیز به عنوان تبلیغ‌کننده در یک شبکه تبلیغ ثبت نام می‌کنند و با جذب کاربران به وب‌سایت خود، در آنجا تبلیغاتی با ارزش بالاتر از یک شبکه تبلیغ دیگر را به نمایش گذاشته و از این طریق کسب درآمد می‌کنند. این منتشرکنندگان، معمولاً با ارائه محتوای ضعیف، کاربران را ترغیب می‌کنند که روی تبلیغات باارزش کلیک نمایند)، ۳) تعداد کلیک‌های انجام شده توسط شبکه رباتی که هزینه ثابت بین آنها سرشکن می‌شود، ۴) درآمد حاصل که به صورت افزایشی به ازای هر کلیک، افزوده می‌گردد. بنابراین، یک منتشرکنندهٔ متقلب تنها به دو طریق می‌تواند نسبت به منتشرکننده‌های درستکار، درآمد خود را افزایش دهد: ۱) افزایش تعداد کلیک روی تبلیغات و ۲) انجام کلیک روی تبلیغات با ارزش بیشتر.

در گام بعد، آنها سیستم پیشنهادی خود را مطرح می‌کنند. سیستم ارائه شده شامل دو

بخش برخط و برون‌خط می‌باشد. بخش برون‌خط به بررسی لاگ‌های جمع آوری شده از کلیک‌ها و تشخیص منتشرکنندگان با رفتار مشکوک می‌پردازد و در بخش برخط از نتایج آن استفاده نموده و با انجام یک کلیک، تصمیم می‌گیرد که کلیک را به عنوان یک کلیک معتبر در نظر گرفته و از اعتبار تبلیغ‌کننده کم کند یا خیر.

به ازای کلیک روی هر تبلیغ، اطلاعاتی مانند منتشرکنندهٔ تبلیغ، کاربر کلیک‌کننده و درآمد حاصل از آن کیک ذخیره می‌گردد. برای هر زوج (کاربر و منتشرکننده)، لگاریتم مجموع درآمدی که کاربر برای وب‌سایت منتشرکننده داشته، محاسبه می‌گردد. سپس به ازای هر منتشرکننده، درآمد حاصل از تمام کاربران آن به صورت نزولی مرتب شده و N مقدار نخست آن در یک بردار نگه‌داری می‌شود (این بردار در بازه‌های زمانی به صورت دوره‌ای به‌روز می‌شود). برای منتشرکننده‌های درستکار، میانگین بردارهای آنان محاسبه و بردار حاصل به عنوان مبنای تشخیص کلیک‌های هرز در نظر گرفته می‌شود. حال به ازای هر منتشرکننده، اختلاف بردار مربوط به آن و بردار مبنا محاسبه می‌گردد و حاصل به عنوان امتیاز منفی برای آن منتشرکننده در نظر گرفته می‌شود. اگر این امتیاز از یک آستانه بیشتر باشد، این منتشرکننده به عنوان متقلب شناخته می‌شود و بردار مربوط به آن در بخش برخط ذخیره می‌شود. در بخش برخط، با آمدن یک کلیک، اگر منتشرکنندهٔ مربوط به آن در بخش برون‌خط به عنوان متقلب شناخته شده باشد و درآمد تولید شده از کاربر برای آن منتشرکننده در محدوده بردار آن منتشرکننده قرار بگیرد، کلیک شمرده نمی‌شود و از اعتبار تبلیغ‌کننده کم نمی‌گردد.

۴-۲-۲ اعتبارسنجی کاربران

بر خلاف روش‌های قبلی که به شناسائی منتشرکنندگان متقلب می‌پرداختند، این دسته از روش‌ها به اعتبارسنجی کاربران می‌پردازند. در [۱۲]، از تعدادی وب‌سایت معتبر و محبوب، برای احراز هویت کاربران استفاده می‌شود، به این ترتیب که هرگاه کاربر فعالیت خاصی در آنها انجام دهد (مثلاً خرید کند)، وب‌سایت مزبور در مرورگر کاربر یک کوپن (توکن رمزنگاری شده) قرار می‌دهد. سپس زمانی‌که کاربر روی یک تلیغ کلیک می‌کند، این کوپن نیز به همراه شناسه

منتشرکننده و شناسه تبلیغ برای شبکه تبلیغ ارسال می‌گردد. هرگاه کاربری دارای این کوپن باشد به این معنی است که او معتبر بوده و کلیک انجام شده از سوی او هم معتبر می‌باشد. این وب‌سایت‌ها وابسته به شبکه‌های تبلیغاتی هستند و برای عملیات درج توکن در مرورگر کاربران و اعتبارسنجی آنها از شبکه‌های تبلیغ کسب درآمد می‌نمایند. شبکه تبلیغ، به محض دریافت یک کلیک، کوپن موجود در آن را بررسی می‌نماید تا اخیراً استفاده نشده باشد (برای شناسائی کلیک‌های تکراری که نوعی از کلیک‌های هرز به حساب می‌آیند). به این ترتیب اگر شبکه تبلیغ، کلیک انجام شده را معتبر دانست از شارژ تبلیغ‌کننده کم می‌نماید. به منظور جلوگیری از ایجاد کوپن‌های جعلی توسط حمله‌کنندگان، شبکه تبلیغ و هریک از وب‌سایت‌های وابسته به آن، یک کلیک مشترک بین خود به اشتراک می‌گذارند و از مکانیزم HMAC برای ایجاد کوپن استفاده می‌کنند.

روش‌های مبتنی بر اعتبارسنجی کاربران، کمتر مورد استفاده قرار می‌گیرند زیرا احراز هویت کاربران به تنهایی نمی‌تواند به معتبر بودن تمامی کلیک‌های انجام گرفته از سوی کامپیوتر کاربر منتج شود زیرا بسیاری از حملات کلیک‌های هرز در قالب بدافزارها روی کامپیوتر کاربران واقعی می‌نشینند و به فعالیت می‌پردازند که با این روش‌ها قابل شناسائی نیستند.

تا اینجا تعدادی از مهم‌ترین روش‌های ارائه شده جهت شناسائی کلیک‌های هرز در شبکه‌های تبلیغاتی مرور شدند. علاوه بر روش‌های عنوان شده، کارهای دیگری نیز در این حوزه مطرح شده‌اند که ما به دلیل اینکه مستقیماً به موضوع این پژوهش مرتبط نیستند، از آنها صرف نظر می‌کنیم [۱۳-۱۵].

۲-۳ شناسائی کلیک‌های هرز در موتورهای جستجو

در این بخش، به مرور کارهایی که صرفاً به مسئله شناسائی کلیک‌های هرز در صفحهٔ نتایج موتورهای جستجو (کلیک روی لینک‌های نتایج و لینک‌های تبلیغاتی) تمرکز کرده‌اند و ما نیز از برخی ایده‌های آنها در این پژوهش استفاده نموده‌ایم، می‌پردازیم. معمولاً این روش‌ها با سه رهیافت مختلف به شناسائی کلیک‌های هرز می‌پردازند: ۱) شناسائی ترافیک‌های غیر نرمال، ۲)

شناسائی کابران غیر نرمال و ۳) شناسائی نشست‌های غیر نرمال. البته تمامی روش‌هایی که در این بخش مرور می‌شوند به صورت برون‌خط روی لاگ فعالیت کاربران (پرس‌وجو و کلیک) موتورهای جستجو مورد ارزیابی و تست قرار گرفته‌اند.

۲-۳-۱ شناسائی کاربران غیر نرمال

گروهی از روش‌ها سعی می‌کنند با شنااسئی کاربران غیر نرمال، کلیک‌های انجام شده توسط آنان را به عنوان هرز در نظر بگیرند. در یکی از این روش‌ها، کنگ و همکارانش [۱۶]، با کمک دسته‌بندی کاربران به دسته‌های "نرمال" و "غیر نرمال" به شناسائی ترافیک غیر نرمال می‌پردازند. در روش ارائه شده توسط آنان، به دلیل فقدان مجموعه داده آموزشی برچسب‌دار، در ابتدا محققان با تدوین مکانیزمی بر اساس وضعیت بار سرورها (زمان‌های اوج ترافیک)، رفتار کاربر نظیر ارسال تعداد زیادی پرس‌وجو در بازه کوتاه، رفتار آدرس‌های IP مانند حجم فعالیت، اقدام به نمایش کپچا به کاربران می‌نمایند. چنانچه کاربری به درستی به کپچا پاسخ دهد در دستهٔ نرمال قرار می‌گیرد ولی اگر به کپچا پاسخ غلط بدهد و یا اصلاً پاسخ ندهد، بر مبنای تعدادی روش شهودی[1] نظیر تعداد کلیک‌ها در یک بازه زمانی، تعداد صفحات بازدید شده در موتور جستجو و تعداد آدرس‌های IP که کاربر به صورت همزمان از آنها به فعالیت می‌پردازد، در مورد دستهٔ آنها نتیجه‌گیری می‌کند. یعنی اگر کاربری به کپچا پاسخ نداد و از آستانه‌های تعریف شده در روش‌های شهودی گفته شده نیز عبور کرد، به عنوان یک کاربر غیر نرمال تلقی می‌شود و با دستهٔ "غیر نرمال" برچسب می‌خورد. در غیر این صورت، در دستهٔ "نا مشخص" قرار می‌گیرد.

آنها پس از تولید مجموعه آموزشی برچسب‌دار، در گام دوم به مدل‌سازی رفتار کاربران در قالب مجموعه‌ای از ویژگی‌ها می‌پردازند. ویژگی‌های به کار رفته در مدل عبارتند از:

- تعداد صفحات مرور شده توسط کاربر
- تعداد کلیک‌های انجام شده روی مجموعه نتایج

[1] Heuristics

- تعداد تأثیراتی[1] که کاربر علاوه بر نتایج جستجو دریافت می‌کند (مانند محتوای چند رسانه‌ای و یا تصاویر که به ازای برخی پرس‌وجوها در کنار نتایج نمایش داده می‌شوند)
- تعداد آدرس‌های IP که کاربر به صورت همزمان از آنها استفاده می‌کند
- تعداد پرس‌وجوهای یکتایی که کاربر در یک نشست ارسال نموده است
- یک فیلد باینری (صفر یا یک) که تحت یکی از شرایط زیر یک می‌شود:
  - آدرس IP مربوط به کاربر جزو لیست سیاه باشد.
  - پرس‌وجوی کاربر حاوی کدهای داخلی موتور جستجو باشد.
  - ترکیب پرس‌وجوی کاربر پیچیده تر از آنی باشد که توسط کاربر تایپ شده باشد.

نهایتاً در گام سوم، یک الگوریتم نیمه نظارتی بر پایه دسته‌بند شبکه بیزین پیشنهاد می‌شود تا ضمن بهبود مجموعه داده آموزش تولید شده بتواند با دقت خوبی به دسته‌بندی کاربران بپردازد. برای ایجاد ساختار شبکه بیزین از ساختار درخت پوشای با حداکثر وزن استفاده می‌شود. همچنین آنها جهت تخمین پارامترها از الگوریتم "بیشینه‌سازی امید ریاضی" با افزودن فاکتور "اعتماد به کپچا" در فرآیند یادگیری پارامترها بهره گرفتند. با افزودن این پارامتر، اگر کاربری به چالش کپچا پاسخ درست دهد، احتمالات پسین مربوط به غیر نرمال بودن او به میزان اعتماد به کپچا کاسته می‌شود و احتمال نرمال بودن وی به همان میزان افزایش می‌یابد.

روش معرفی شده در این پایان‌نامه نیز مبتنی بر دسته‌بندی است اما با الگوریتم فوق دارای تفاوت های اساسی می‌باشد که مهم‌ترین آنها عبارتند از:

۱) روش تولید مجموعه آموزشی برچسب‌دار در این پژوهش مستقل از تکنیک کپچا می‌باشد.

۲) ما از ویژگی‌هایی در سه سطح نشست، کاربر و آدرس IP برای مدل‌سازی داده‌ها استفاده می‌کنیم در حالی که در مقاله مذکور فقط از ویژگی‌های سطح کاربر استفاده شده است. البته،

[1] Impression

چون عمده ویژگی‌های عنوان شده در این تحقیق، از نشست جاری کاربر محاسبه می‌شود، روش پیشنهادی ما در دسته روش‌های بخش ۲-۳-۳ قرار می‌گیرد.

۳) در این تحقیق به جای دسته‌بند شبکه بیزین از دسته‌بند K-نزدیک‌ترین همسایه استفاده شده است و در کنار آن، مکانیزمی برای بهبود داده‌های آموزشی در نظر گرفته شده است.

۴) روش پیشنهادی ما، کاملاً قابلیت به کارگیری به صورت برخط را دارا می‌باشد در حالی که روش فوق صرفاً به صورت برون‌خط قابل استفاده است.

۲-۳-۲ شناسائی ترافیک‌های غیر نرمال

بخشی از حمله کنندگان با ارسال ترافیک توزیع شده و با نرخ پایین سعی در پنهان کردن رفتار خود می‌نمایند. تشخیص این نوع از حملات به کمک روش‌های مبتنی بر شناسائی کاربران غیر نرمال چندان امکان پذیر نیست بلکه یک موتور جستجو می‌بایست تمام ترافیک دریافتی از مجموعه کاربران را به صورت کلی بررسی نماید. روش ارائه شده در [۱۷]، نمونه‌ای از این دسته روش‌ها است که قادر به شنااسئی شبکه‌های رباتی توزیع شده با حجم ترافیک کم می‌باشد. آنها از لاگ جستجو و کلیک کاربران برای بررسی روش خود استفاده نمودند. روش پیشنهاد شده توسط این محققان در دو گام انجام می‌گیرد: شناسائی پرس‌وجوهای مشکوک و شناسائی شبکه‌های رباتی.

در گام نخست، به ازای هر یک از پرس‌وجوهای کاربران، دو هیستوگرام از لینک‌های موجود در صفحه نتایج (که معمولاً ۱۰ مورد می‌باشد) ساخته می‌شود که اولی بر اساس فرکانس کلیک‌های اجام شده در ماه جاری و دیگری مربوط به لاگ‌های قبلی (مثلاً ماه قبل) می‌باشد. اگر دو هیستوگرام تفات نسبتاً زیادی با یکدگیر داشته باشند آن پرس‌وجو به مجموعه "پرس‌وجوهای مشکوک" افزوده می‌گردد. همچنین پرس‌وجوهایی که در بازه زمانی اخیر به دفعات زیادی ارسال شده‌اند نیز به مجموعه مشکوک اضافه می‌شوند. این مجموعه پرس‌وجوها تنها شامل فعالیت‌های غیر نرمال نمی‌شوند بلکه پرس‌وجوهایی که به صورت مقطعی در میان کاربران محبوب می‌گردند را نیز (مثلاً پرس‌وجوی "عید" در نزدیکی ایام نوروز) شامل می‌شود که برای تفکیک این دو مورد، در

گام بعد از روش ماتریسی استفاده می‌شود.

در گام دوم، به ازای هر یک از پرس‌وجوهای موجود در مجموعه مشکوک، کلیه کاربرانی که آن پرس‌وجو را ارسال نموده‌اند و همچنین سایر پرس‌وجوهایی که توسط آنان ارسال شده، استخراج می‌شود. سپس ماتریسی از پرس‌وجوها-کاربران ایجاد می‌شود که سطرها بیانگر پرس‌وجوهای ارسال شده و ستون‌ها معرف کاربران هستند. مقدار هر درایه نیز تعداد دفعات ارسال پرس‌وجو توسط کاربر را نشان می‌دهد. مشاهده می‌شود که در ترافیک‌های تولید شده از ربات‌هاُ ستون‌های ماتریس شباهت زیادی به هم دارند اما در ماتریس مربوط به پرس‌وجوهای محبوب، کاربران (ستون‌ها) عموماً تنها در همان یک پرس‌وجو با هم مشترک هستند.

در برخی موارد، ربات‌هاتنها یک یا تعداد محدودی پرس‌وجو ارسال می‌کنند، یعنی تعداد سطرهای ماتریس بسیار کم می‌شود. برای شناسائی این ربات‌ها پیشنهاد شده است که نسبت ترافیک ارسال شده برای تنها آن پرس‌وجو به کل ترافیک ارسال شده از سوی کاربرانی که این پرس‌وجو را ارسال نموده‌اند، محاسبه شود. در صورتی که این نسبتُ بیشتر از یک آستانه تعریف شده باشد، به عنوان گروه رباتی تشخیص داده می‌شود.

به عقیده محققان، در بسیاری از شبکه‌های رباتی به دلیل مدیریت متمرکزُ معمولاْ مجموعه مشابه یا مشترکی از پرس‌وجوها در میان ربات‌های آنها استفاده می‌شود اما هر ربات، با افزودن تعداد زیادی پرس‌وجوی تصادفی و یا پرس‌وجوهای محبوب در میان کاربران سعی می:ند فعالیت خود را متفاوت از دیگر ربات‌ها نشان دهد. این نوع حملات به دلیل ترکیب با رفتار نرمال کاربران، ممکن است با روش گفته شده قابل شناسائی نباشد. برای تشخیص این موارد، از الگوریتم PCA استفاده می‌شود. الگوریتم PCA یک روش آماری جهت کاهش بعد بوده که قادر است داده‌هایی با ابعاد بالا را به داده‌هایی با ابعاد کوچکتر نگاشت نماید. بنابراین، با اعمال الگوریتم PCA روی بردار کاربران (ستون‌ها در ماتریس پرس‌وجوها-کاربران) می‌توان از میان پرس‌وجوهای نرمال کاربران، پرس‌وجوهای همبسته را تشخیص داد. به این ترتیب بعد از به دست آوردن بزرگ‌ترین مؤلفه اصلی از الگوریتم PCA، داده‌ها به آن فضا نگاشت می‌شوند. حال اختلاف بردار هر کاربر در فضای نگاشت شده با داده‌های اصلی (پیش از نگاشت) محاسبه می‌شود. چنانچه این اختلاف بسیار کوچک باشد،

به این معنی است که کاهش بعد نتوانسته تغییر چندانی در آنها ایجاد نماید که این نشان از فعالیت رباتی می‌دهد. بنابراین در نهایت، K کاربری که بردارهای آنان کمترین اختلاف با حالت نگاشت شده خود در فضای جدید دارند به عنوان یک گروه رباتی در نظر گرفته می‌شوند.

۲-۳-۳ شناسائی نشست‌های غیر نرمال

دستهٔ دیگری از روش‌های تشخیص کلیک‌های هرز از طریق شناسائی نشست‌های غیر نرمال کاربران اقدام می‌کنند. زمانی‌که یک کاربر وارد موتور جستجو می‌شود، یک شناسه یکتا به عنوان "شناسه نشست" به وی تخصیص داده می‌شود. این شناسه بعد از گذشت یک محدودیت زمانی (معمولاْ ۳۰ دقیقه) از فعال نبودن کاربر منقضی می‌شود و وقتی آن کاربر مجدداً به سایت برگشت، یک شناسه نشست جدید به او داده می‌شود. در روش‌های ارائه شده در این گروه، مجموعه فعالیت‌های کاربر در یک نشست مورد بررسی قرار گرفته و در مورد غیر نرمال بودن آن تصمیم‌گیری می‌شود. روش‌های جدیدتر، معمولاْ در این دسته قرار می‌گیرند که در اینجا به مرور دو نمونه تحقیق از این دسته می‌پردازیم.

در [۱۸]، محققان با هدف تشخیص نشست‌های متعارف و غیر متعارف، روشی مبتنی بر مدل‌سازی نشست‌های کاربران به صورت دنباله‌ای از فعالیتآ مطرح کردند. آنها یک نشست را متعارف می‌دانند اگر دنباله فعالیت‌های انجام شده در آن از یک توالی منطقی پیروی نماید و در غیر این صورت آن را غیر نرمال در نظر می‌گیرند. روش مطرح شده شامل سه بخش می‌شود: ۱) مدل‌سازی نشست‌های کاربران با استفاده از زنجیره مارکوف، ۲) ایجاد مدل هفت بعدی از رفتار نشست‌ها، ۳) شناسائی نشست‌های غیر نرمال.

در گام نخست، هر نشست از کاربران به صورت دنباله‌ای از جفت‌های (نوع فعالیت و شماره صفحه) بیان می‌شود. "نوع فعالیت" کاربر می‌تواند شامل یکی از ۵ مورد زیر باشد:

- ارسال پرس‌وجو که با P نمایش داده می‌شود.
- کلیک روی نتایج جستجو که با W نمایش داده می‌شود.
- کلیک روی لینک‌های تبلیغاتی موجود در صفحه نتایج که با O نمایش داده

می‌شود.

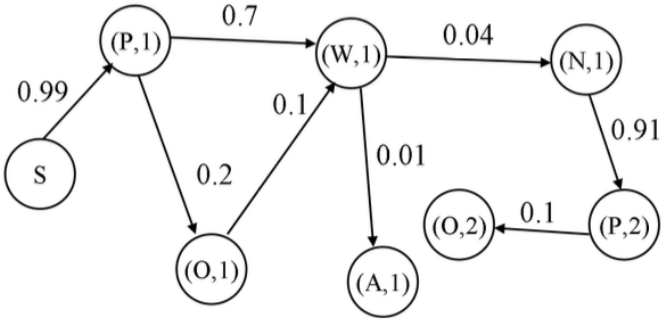
- کلیک روی شماره صفحات دیگر از مجموعه نتایج که با N نمایش داده می‌شود.
- سایر کلیک‌های انجام شده در صفحه که با A نشان داده می‌شود.

منظور از "شماره صفحه"، شماره صفحه‌ای از مجموعه نتایج است که کاربر در حال حاضر (پیش از فعالیت بعدی) در آن حضور داشته است. سپس مجموع زوج‌ها به صورت زنجیره مارکوف مدل می‌شوند. فضای حالت در این زنجیره مارکوف، شامل تمام زوج فعالیت‌هایی است که کاربر در آن نشست انجام داده که با احتساب وضعیت شروع (S) در هر نشست، مجموعه فضای حالت به صورت $\{S\} \cup \{\{P.W.N.O.A\} * N\}$ بیان می‌شود. آنگاه احتمال انتقال از وضعیت i به j به صورت

$$\Pr(i.j) = \frac{Q_{i.j}}{Q_i} \qquad (۱-۲)$$

محاسبه می شود که در آن برابر با تعداد انتقالات از حالت i به j در کل مجموعه نشست‌ها و برابر با تعداد کل انتقالاتی است که از وضعیت i شروع می‌شوند.

بخشی از یک زنجیره مارکوف در شکل ۲-۲ نشان داده شده است.



شکل ۲-۲: نمونه‌ای از زنجیره مارکوف با احتمالات انتقالی بین حالات مختلف [۱۸]

سپس به ازای هر نشست، امتیازی محاسبه می‌شود که میزان نامتعارف بود آن نشست را بیان می‌کند. این امتیاز، از حاصل‌ضرب احتمال انتقال از هر فعالیت به فعالیت بعدی در آن نشست (φ) به دست می‌آید. نشست‌های طولانی‌تر به دلیل ضرب احتمالات در یکدیگر، حاصل کوچکتری خواهند داشت لذا امتیاز نهایی به صورت زیر محاسبه می‌شود:

$$MLH_{avg} = \frac{\ln(\varphi)}{S} \quad (۲-۲)$$

که در آن S تعداد فعالیت‌های انجام شده در آن نشست می‌باشد. این مقدار به عنوان یک ویژگی در مدل هفت بعدی پیشنهاد شده در گام بعد، استفاده می‌شود.

در گام دوم، برای شناسائی تعداد بیشتری از نشست‌های نامتعارف، هر یک از نشست‌ها به صورت یک مدل هفت بعدی مدل‌سازی می‌شوند. ویژگی‌های مطرح شده در این مدل عبارتند از: ۱) احتمال زنجره مارکوف (MLH_{avg})، ۲) تعداد فعالیت‌های انجام شده در نشست، ۳) نسبت تعداد پرس‌وجوهای ارسالی به کل فعالیت‌ها در نشست، ۴) نسبت تعداد کلیک‌های انجام شده به کل فعالیت‌ها در نشست، ۵) نسبت تعداد کلیک‌های روی لینک‌های تبلیغاتی به کل فعالیت‌ها، ۶) نسبت تعداد کلیک روی شماره صفحات دیگر به کل فعالیت‌ها، ۷) نسبت سایر کلیک‌های انجام شده در نشست به کل فعالیت‌ها.

پس از مدل‌سازی تمامی نشست‌ها، در گام سوم فاصله هر نشست از میانگین نشست‌ها با استفاده از معیار فاصله ماهالانوبیس محاسبه می‌شود:

$$d = \sqrt{(q-\mu)\Sigma^{-1}(q-\mu)^T} \quad (۳-۲)$$

که در این رابطه، q برداری از یک نشست مدل شده، μ بردار میانگین همه نشست‌ها و Σ ماتریس کواریانس می‌باشد. فواصل زیاد به معنای نشست غیر متعارف و فواصل کم به معنای متعارف بودن آن نشست و فعالیت‌های انجام گرفته در آن در نظر گرفته می‌شود.

در [۱۹] نیز از نشست‌های کاربران برای شناسائی کلیک‌های هرز استفاده می‌شود. در پژوهش قبلی، معنا و مفهوم فعالیت‌های کاربر در نظر گرفته نمی‌شد مثلاً ارسال متوای دو پرس‌وجوی مختلف یا دو پرس‌وجوی یکسان، به یک صورت در نظر گرفته می‌شوند. لذا در این مقاله پیشنهاد می‌شود که نشست‌های کاربران به صورت توالی‌های سه تایی از (نوع فعالیت، هدف فعالیت و اختلاف زمانی فعالیت جاری از فعالیت ماقبل خود) مدل شود. ۶ نوع فعالیت برای کاربران تعریف می‌شود:

- Q_i: ارسال پرس‌وجو که i برای تمایز بین پرس‌وجوهای مختلف به کار می‌رود.
- W_i: کلیک روی نتایج جستجو که i به معنای کلیک روی لینک‌های متفاوت می‌باشد.

- O_i: کلیک روی نتایج تبلیغاتی مختلف
- N: کلیک روی یک شماره صفحه دیگر از مجموعه نتایج
- T: پیمایش صفحه نتایج
- A_i: کلیک روی سایر لینک‌های موجود در صفحه نظیر کلیک روی برگه "ویدئو" یا "تصویر"

همانطور که دیده می‌شود، به برخی فعالیت‌ها یک شناسه یکتا داده می‌شود که هدف آن فعالیت را مشخص می‌کند مثلاْ اگر کاربر دو پرس‌وجوی یکسان را در دو نوبت وارد نماید، هر دو یک شناسه می‌گیرند اما پرس‌وجوهای مختلف شناسه‌های متفاوتی می‌گیرند. جزء سوم نیز اختلاف زمانی هر فعالیت را نسبت به فعالیت قبلی‌اش در نشست بیان می‌کند که برای آن یکی از مقادیر $T = \{0.1.2.3\}$ لحاظ می‌شود؛ اگر اختلاف زمانی دو فعالیت متوالی برابر با صفر باشد مقدار ۰، اگر کمتر از ۱۰ ثانیه باشد مقدار ۱، اگر بین ۱۰ تا ۳۰ ثانیه باشد مقدار ۲ و برای اختلافات بیش از ۳۰ ثانیه عدد ۳ در نظر گرفته می‌شود.

در گام بعد، پژوهشگران از طریق آنالیز نشست‌ها به ایجاد مجموعه اولیه از نشست‌های غیر نرمال اقدام کرده و تعداد محدودی از حالات مختلف را که بیانگر رفتار غیر نرمال هستند، استخراج می‌نمایند. این حالات عبارتند از:

- (QA_i) *: ارسال پرس‌وجوهای مختلف اما کلیک روی لینک‌های با یک دامنه خاص به صورت تکراری در فواصل زمانی کوتاه
- (Q_iT) *: ارسال یک پرس‌وجوی مشخص و پیمایش صفح نتایج به صورت مداوم
- (Q_i) *: ارسال یک پرس‌وجوی خاصل به صورت تکراری
- $Q(W_i)$ *: ارسال یک پرس‌وجو و انجام کلیک‌های تکراری روی یک لینک خاص
- $Q(A_i)$ *: ارسال یک پرس‌وجو و سپس انجام کلیک‌های پشت سر هم روی لینک‌هایی با یک دامنه مشخص

اگر بیش از ۵۰٪ از فعالیت‌های یک نشست با یکی از حالت‌های فوق تطبیق یابد، آن نشست به مجموعه نشستّای غیر نرمال افزوده می‌گردد. به این ترتیب مجموعه‌ای از نشستّای غر

نرمال ایجاد می‌شود.

سپس نویسندگان مقاله، با این نظریه که "اگر کاربری تعداد قابل ملاحظه‌ای نشست غیر نرمال داشته باشد، به احتمال زیاد سایر نشست‌های او نیز غیر نرمال هستند"، الگوریتم گراف دو بخشی "کاربر-نشست" را معرفی نمودند. در این الگوریتم، به نشست‌های غیر نرمال شناسائی شده در گام قبل، امتیاز "یک" و به سایر نشست‌ها امتیاز "صفر" داده می‌شود. آنگاه به هر کاربر امتیازی برابر با میانگین وزن‌دار از امتیاز تمامی نشست‌هایی که تاکنون داشته است، تخصیص داده می‌شود. سپس مجدداً، امتیاز هر نشست به صورت میانگین وزن‌دار امتیاز کاربرانی که آن نشست را داشته‌اند، به روز رسانی می‌شود. فرآیند به روز رسانی امتیاز کاربران و نشست‌ها ادامه می‌یابد تا زمان که اختلاف امتیازها در دو دور متوالی ناچیز باشد. در پایان، نشست‌هایی که امتیاز آنها بیشتر از یک آستانه تعریف شده باشد به عنوان نشست غیر نرمال در نظر گرفته می‌شود و به عنوان ورودی به مرحله بعد جهت شناسائی نشست‌های غر نرمال بیشتر داده می‌شوند.

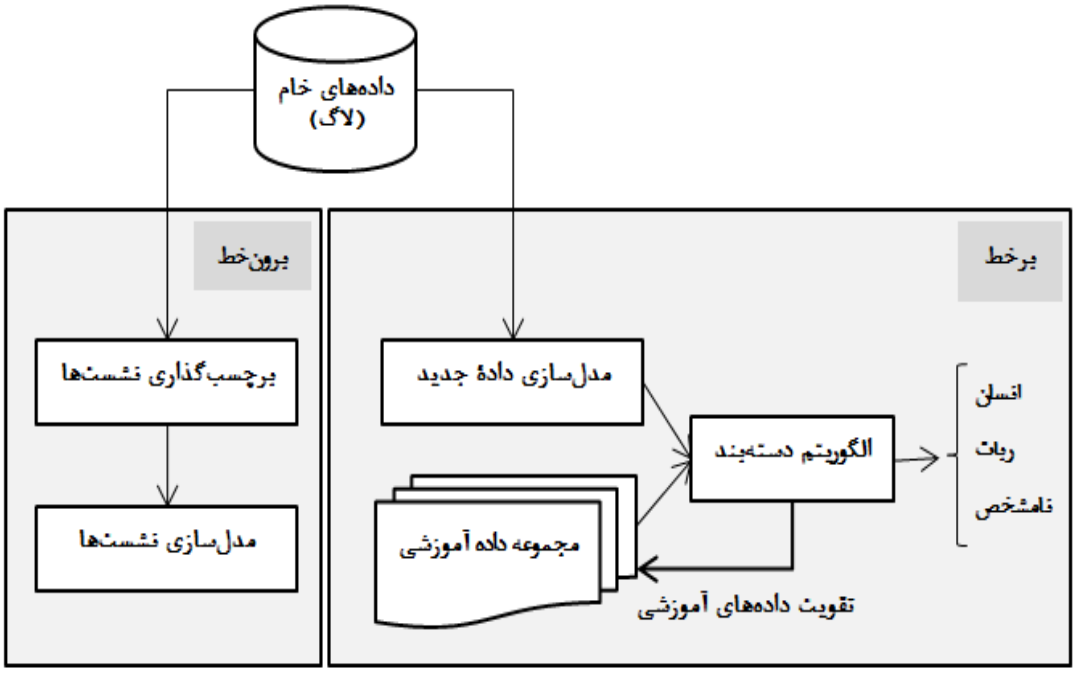
در مرحله آخر، با این استدلال که "اگر تعدادی از نشست‌های غیر نرمال داریا یک الگوی مشترک باشند به احتمال زیاد سایر نشست‌هایی که از این الگو پیروی می‌نمایند نیز غیر نرمال خواهند بود"، الگوریتم گراف دو بخشی "الگو-نشست" پیشنهاد داده می‌شود. برای این منظور ابتدا با استفاده از الگوریتم PrefixSpan، تمامی الگوهای موجود در نشست‌های غیر نرمال شناسائی شده در گام قبل استخراج می‌شوند. سپس مشابه الگوریتم گراف "کاربر-نشست"، به ازای تمامی نشست‌ها، الگوهای آنها استخراج و امتیاز الگوها و نشست‌ها محاسبه و به روز رسانی می‌شود. در پایان، نشست‌هایی که امتیاز آنها فراتر از آستانه در نظر گرفته شده باشند، به عنوان نشست غیر نرمال در نظر گرفته می‌شود و تمامی کلیک‌های انجام شده در آنها به عنوان کلیک هرز در نظر گرفته می‌شود.

فصل سوم: روش پیشنهادی

۳-۱ مقدمه

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

شمای کلی از سیستم پیشنهاد شده در شکل ۳-۱ نشان داده شده است.



شکل ۳-۱: شمای کلی از سیستم پیشنهادی

ما در این تحقیق، 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 می‌کنیم. پس از تولید مجموعه داده آموزشی،

وارد فاز دسته‌بندی می‌شویم. در این فاز، الگوریتم دسته‌بندی پیشنهادی که گونه‌ای تغییر یافته از الگوریتم K-نزدیک‌ترین همسایه می‌باشد را معرفی می‌کنیم. این الگوریتم در دو نسخه دو کلاسه و تک کلاسه مطرح می‌شود. در انتها، تکنیکی جهت تقویت مجموعه داده آموزشی و افزایش دقت دسته‌بندی ارائه می‌دهیم.

۳-۲ مدل‌سازی داده‌ها

داده‌های استفاده شده در این پژوهش، لاگ فعالیت کاربران در یکی از پربازدیدترین موتورهای جستجوی فارسی (به آدرس parsijoo.ir) می‌باشد. هر سطر 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. برای محاسبه ویژگی‌ها، درخواست‌های کاربران به ترتیب زمان واقعی‌شان پیمایش می‌شوند. به ازای هر درخواست مجموعه‌ای از ویژگی‌ها محاسبه می‌شود و سپس الگوریتم دسته‌بندی پیشنهادی روی آن اعمال و در مورد نرمال و غیر نرمال بودن آن تصمیم‌گیری می‌شود.

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

۳-۲-۱ ویژگی‌های سطح نشست

برای محاسبه ویژگی‌های سطح نشست، ما فعالیت‌های کاربر را در هر نشست مورد بررسی قرار می‌دهیم. همانطور که قبلاً هم اشاره شد، زمانی‌که یک کاربر وارد موتور جستجو می‌شود، یک شناسه یکتا به عنوان "شناسه نشست" به وی تخصیص داده می‌شود. این شناسه بعد از گذشت یک محدودیت زمانی از فعال نبودن کاربر (معمولاً ۳۰ دقیقه) منقضی می‌شود و وقتی آن کاربر مجدداً به سایت برگشت، یک شناسه نشست جدید به ی 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 پژوهش، ما به دلیل عدم دسترسی به سایر اقدامات کاربر، تنها سه نوع فعالیت زیر 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 سرویس (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"، "صفه 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 [۱۸] مطرح شده و مقدار آن همانطور که دربخش ۲-۳-۳ گفته شد، پس از مدل‌سازی فعالیت‌های کاربر به صورت زنجیره مارکوف، از حاصل‌ضرب احتمال انتقال از یک وضعیت به وضعیت بعدی به دست می‌آید.
- تعداد کل فعالیت‌های کاربر در نشست بر حسب (Q، W، N)
- 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 یکتای کلیک شده به کل لینک‌های کلیک شده
- 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 صفحات دیگر به مدت زمان فعالیت نشست

هفت ویژگی آخر برای اولین بار در این پژوهش پیشنهاد شده‌اند و مابقی در [۱۸] و [۲۰] به کار رفته‌اند. مقادیر تمام ویژگی‌های فوق و همینطور ویژگی‌هایی که در ادامه خواهند آمد به بازه صفر تا یک نرمال‌سازی می‌شوند.

۳-۲-۲ ویژگی‌های سطح کاربر

مشابه "شناسه نشست"، هر کاربر یک "شناسه کاربری" نیز دارد که در مرورگر او و همچنین موتور جستجو ذخیره می‌شود. این شناسه برخلاف شناسه نشست که پس از یک محدودیت زمانی منقضی می‌شود، تا زمانی‌که توسط خود کاربر از مرورگرش حذف نشود، پا برجا باقی می‌ماند.

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.

۳-۲-۳ ویژگی‌های سطح آدرس 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 در نظر گرفتیم، عبارتند از:

- آدرس IP اینترانت یا اینترنت (۰ یا ۱): کاربرانی که از سازمان‌ها و ادارات داخل

کشوری هستند، دارای آدرس IP اینترانت می‌باشند.

- تعداد کل فعالیت‌های هر آدرس IP
- تعداد پرس‌وجوهای ارسال شده
- تعداد کل کلیک‌های انجام شده روی شماره صفحات
- نسبت لینک‌های یکتای کلیک شده به کل لینک‌های کلیک شده
- نسبت دامنه‌های یکتای کلیک‌شده به کل لینک‌های کلیک شده
- نسبت کوکی‌های تخصیص یافته به کل فعالیت‌های هر آدرس IP

هر نشست ممکن است با یک یا چند آدرس IP همراه باشد، بنابراین ویژگی‌های فوق به صورت میانگین روی تمامی آدرس‌های آن نشست محاسبه می‌گردد. همچنین لازم است ذکر گردد که ما اطلاعات نگه‌داری شده برای هر آدرس IP را در بازه‌های زمانی متناوب، هرگاه یک آدرس IP بیشتر از ۳۰ دقیقه فعالیت نداشته باشد، اطلاعات نگه‌داری شده از آن را صفر می‌نماییم که چون این اتفاق معمولاً در شب‌ها رخ می‌دهد، لذا ما این کار را برای تمامی آدرس‌های IP در ساعت ۱۲ شب انجام می‌دهیم.

۳-۳ تولید مجموعه داده آموزشی اولیه

حجم بالای لاگ موتورهای جستجو، برچسب‌گذاری دستی آنها را به عنوان "انسان یا ربات" جهت تولید مجموعه داده آموزشی 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 [۲۱] ایده 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، برای تولید مجموعه آموزشی، پیشنهاد نمایش کپچا تنها به بخش کوچکی از کاربران را مطرح کردند [۱۶]. در این روش که در بخش ۲-۳-۱ نیز به معرفی آن پرداختیم، تنها به کاربرانی که از آستانه‌های تعریف شده برای چند پارامتر ساده مانند وضعیت بار سیستم و تعداد پرس‌وجوهای ارسالی فراتر رفته‌اند، 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 اکثراً نیز به آن جواب نداده‌اند، بنابراین می‌توان استنباط نمود که این درخواست‌ها از سوی برنامه‌های رباتی بوده که قادر به حال کپچا نیستند.

در این تحقیق، ما برای تولید مجموعه داده آموزشی برچسب‌دار اولیه، از روش مطرح شده در [۱۸] استفاده می‌کنیم. این روش، از از مجموعه روش‌هایی است که از طریق شناسائی نشست‌های غیر نرمال، به تشخیص کلیک‌های هرز می‌پردازد و ما در بخش ۲-۳-۳ به تشریح آن پرداختیم. همانطور که قبلاً اشاره شد، این روش بر پایه یک مدل هفت بعدی استوار است اما ما در اینجا به دلیل عدم دسترسی به کلیک‌های انجام گرفته روی لینک‌های تبلیغات و همچنین سایر کلیک‌های انجام شده در صفحه نتایج (به غیر از کلیک‌های روی نتایج و شماره صفحات)، تنها از پنج ویژگی زیر استفاده نموده‌ایم:

- میانگین احتمال زنجیره مارکوف برای توالی رفتار کاربر در نشست
- تعداد کل فعالیت‌های کاربر در نشست (شامل ارسال پرس‌وجو، کلیک روی نتایج، کلیک روی صفحات بعدی)
- نسبت تعداد دفعات ارسال پرس‌وجو به کل فعالیت‌ها
- نسبت تعداد کلیک‌های روی نتایج به کل فعالیت‌ها
- نسبت تعداد کلیک‌های روی شماره صفحات به کل فعالیت‌ها

بنابراین، بعد از مدل‌سازی نشست‌های کاربران به کمک این مدل پنج بعدی و محاسبه معیار فاصله ماهالانوبیس به ازای آنها، ۱٪ از نشست‌هایی که بیشترین فاصله را داشته‌اند به عنوان نشست نرمال در نظر گرفتیم. در ادامه، هر یک از این نشست‌ها مطابق بخش ۳-۲ مدل‌سازی شده

و سپس نشست‌های غیر نرمال با برچسب "مثبت" یا "یک" و نشست‌های نرمال با برچسب "منفی" یا "صفر" به مجموعه داده آموزشی اضافه می‌گردند.

این مجموعه داده آموزشی، به عنوان یک مجموعه اولیه برای شروع فرآیند دسته‌بندی استفاده می‌گردد اما با روشی که در ادامه خواهیم گفت به تدریج، به بهبود کیفیت مجموعه آموزشی کمک کرده و می‌توانیم با دقت بالاتری به دسته‌بندی نشست‌ها اقدام نماییم.

۳-۴ الگوریتم دسته‌بند پیشنهادی

در روش دسته‌بندی مطرح شده ما از الگوریتم K-نزدیک‌ترین همسایه (KNN) به عنوان الگوریتم پایه استفاده کرده و سپس بنا به ضرورت، تغییراتی در آن لحاظ می‌نماییم. بنابراین لازم است ابتندا به معرفی آن بپردازیم.

الگوریتم K-نزدیک‌ترین همسایه [۲۲]، 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 داده جدید (داده آزمایشی)، آن را با کلیه نمونه‌های موجود در مجموعه داده آموزشی مقایسه و K-نزدیک‌ترین نمونه به آن را استخراج کرده و بر اساس برتری دسته یا برچسب مربوط به آنها، در مورد دسته داده آزمایشی مزبور تصمیم‌گیری می‌نماید.

در یک موتور جستجو، ما با کاربرانی با تنوع رفتاری بالا مواجه هستیم، لذا برای دسته‌بندی دقیق، ناگزیر به نگه‌داری تعداد زیادی داده آموزشی هستیم. از اینرو، 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 محاسبه فاصله داده جدید با تمام نمونه‌های آموزشی به وجود می‌آید. هرچه مجموع داده آموزشی بزرگتر شود، دو مشکل گفته شده بیشتر خود را نشان می‌دهند. ما سعی می‌کنیم در این پایان‌نامه با ارائه ایده‌هایی بر دو مشکل فوق فائق آییم و الگوریتم را برای کاربردمان مناسب‌سازی نماییم.

در این تحقیق، ما دو گونه از الگوریتم KNN را مورد استفاده قرار می‌دهیم: دو کلاسه و تک

کلاسه که به ترتیب در دو بخش آتی به مرور آنها و معرفی تغییرات اعمال شده در آنها برای روش پیشنهاد شده می‌پردازیم.

۳-۴-۱ الگوریتم K-نزدیک‌ترین همسایه دو کلاسه

الگوریتم دسته‌بندی K-نزدیک‌ترین همسایه دو کلاسه، از نمونه‌های آموزشی "مثبت" و "منفی" در مجموعه آموزشی خود استفاده می‌کند.

در این روش، اگر مجموعه داده‌های آموزشی را به صورت $T = \{(x_i . y_i)\}_{i=1}^{N}$ در نظر بگیریم به نحوی که $x_i \in \mathbb{R}^m$ نمونه آموزشی i-اُم در فضای ویژگی بعدی و y_i برچسب متناظر با آن باشد، برچسب نمونه آزمایشی x' در دو گام تعیین می‌شود:

در گام نخست، فاصله نمونه آزمایشی از تمام نمونهٔای موجود در مجموعه آموزشی محاسبه می‌گردد. برای این منظرو از معیار فاصله اقلیدسی که متداول‌ترین معیار فاصله است، استفاده می‌گردد:

$$d(x' . x_i) = \|x' - x_i\|_{L_2} \qquad (۱-۳)$$

که در این رابطه، منظر از نُرم بردار اختلافات، مجذور مجموع مربعات مقادیر آن بردار می‌باشد. K نمونه آموزشی که کمتری فاصله را تا داده آزمایشی داشته باشند در مجموعه همسایگی آن قرار می‌گیرند که این مجموعه همسایگی را با NN نمایش می‌دهیم.

در گام بعد، از برچسب دسته K نزدیک‌ترین همسایه،برای پیش‌بینی دسته نمونه آزمایشی استفاده می‌شود. به این صورت که بین K همسایه رأی‌گیری شده و دسته‌ای که بیشتری دفعات دیده شدن را در بین این K نمونه دارا است، به عنوان دسته نمونه آزمایشی در نظر گرفته خواهد شد:

$$y' = \arg\max_{y} \sum_{(x_i^{NN} . y_i^{NN})} I(y = y_i^{NN}) \qquad (۲-۳)$$

که نمونه آموزشی i-اُم در مجموعه همسایگی NN و برچسب پیش‌بینی شده برای نمونه آزمایشی می‌باشد. رابطه نیز به صورت زیر تعریف می‌شود:

$$I(y) = \begin{cases} 1 & y = 1 \\ 0 & y = -1 \end{cases} \quad (۳-۳)$$

در رابطهٔ (۳-۲)، هریک از نمونه‌های موجود در مجموعه همسایگی سهم یکسانی در تعیین دسته داده آزمایشی دارند در حالی که می‌دانیم بهتر است همسایه‌های نزدیک‌تر، مشارکت بیشتری در تعیین برچسب داده آزمایشی داشته باشند و همسایه‌های دورتر سهم کمتر. برای منظور، معمولاً از یک مکانیزم وزن‌دهی استفاده می‌شود. یکی از مرسوم‌ترین فاکتورهای وزن‌دهی، ضریب $1/d^2$ می‌باشد که d بیانگر فاصله داده آزمایشی تا آن همسایه می‌باشد. بنابراین، با اعمال این ضریب، رابطهٔ (۳-۲) به صورت زیر بازنویسی می‌شود:

$$y' = \arg\max_{y} \sum_{(x_i^{NN}.y_i^{NN})} \frac{1}{d^2(x'.x_i^{NN})} * I(y = y_i^{NN}) \quad (۴-۳)$$

علاوه بر این، ما به هر داده آموزشی یک فیلد شمارنده نیز اضافه می‌نماییم، یعنی مجموعه داده آموزشی به صورت $T = \{(x_i.c_i.y_i)\}_{i=1}^{N}$ تبدیل می‌شود. هدف از افزودن این فیلد آن است که هر داده آموزشی بتواند به نمایندگی از چندین نقطه در مجموعه داده آموزشی حضور داشته باشد. بنابراین می‌توان بنا بر ظرفیت حافظه، یک محدودیت برای تعداد داده‌های آموزشی در نظر گرفت و از این طریق به مشکلات حافظه مصرفی و حم محاسبات فائق شد. از طرف دیگر، باید کیفیت داده‌های آموزشی را نیز افزایش داد تا دقت فرآیند دسته‌بندی بیشتر شود. برای این منظور، رابطه (۳-۴) را به فرم زیر تبدیل می‌نماییم:

$$score(x') = \frac{\sum_{i=1}^{K} \frac{c_i}{d^2(x'.x_i^{NN})} * y_i^{NN}}{\sum_{i=1}^{K} \frac{c_i}{d^2(x'.x_i^{NN})}} \quad (۵-۳)$$

بنابراین، به ازای هر داده آزمایشی، امتیاز دسته‌بندی مربوط به آن طبق رابطه (۳-۵) محاسبه می‌گردد. مقدار این امتیاز در بازه [۱-۰] قرار می‌گیرد. هرچه این مقدار به صفر نزدیک باشد، بیانگر رفتار نرمال بوده و هرچه به یک نزدیک شود، معرف رفتاری غیر نرمال (رفتاری رباتی) خواهد بود. نحوه به روز رسانی مقادیر در بخش ۳-۵ تشریح می‌شود.

در ادامه، ما دو حد آستانه برای این امتیاز تعریف می‌کنیم: آستانه انسانی و آستانه رباتی. اگر امتیاز محاسبه شده، کمتر از آستانه انسانی باشد، داده آزمایشی به عنوان "انسان" یا نمونه "منفی" دسته‌بندی می‌شود و اگر امتیاز محاسبه شده بیشتر از آستانه رباتی باشد، به عنوان

"ربات" یا نمونه "مثبت" برچسب زده می‌شود. در صورتی که امتیاز به دست آمده، بین دو آستانه قرار بگیرد، نمونه آزمایشی برچسب "نامشخص" دریافت می‌کند که در این حالت، ما منتظر فعالیت‌های بیشتر از کاربر می‌مانیم. نحوه انتخاب دو آستانه در فصل بعد تشریح می‌شود.

۳-۴-۲ الگوریتم K-نزدیک‌ترین همسایه تک کلاسه

در روش دسته‌بندی تک کلاسه بر خلاف روش دو کلاسه، تنها نمونه‌های "مثبت" در مجموعه آموزشی نگه‌داری شده و برای عملیات دسته‌بندی مورد استفاده قرار می‌گیرند. مهم‌ترین مزیت این روش در مقایسه با روش دو کلاسه، کاهش حجم داده‌های آموزشی و در نتیجه کاهش حجم محاسبات پردازشی در زمان دسته‌بندی می‌باشد. در این بخش نیز 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 به صورت $T = \{x_i\}_{i=1}^{N}$ در نظر بگیریم به نحوی که x_i نمونه آموزشی i-اُم در فضای m بعدی باشد، برچسب نمونه آزمایشی x' در دو گام مشخص می‌گردد:

گام اول، عیناً مانند روش دو کلاسه است که در آن فاصله نمونه آزمایشی از تمام نمونه‌های موجود در مجموعه آموزشی محاسبه می‌گردد و K نمونه آموزشی که کمترین فاصله را تا داده آزمایشی داشته باشند در مجموعه همسایگی آن قرار می‌گیرند.

اما در گام دوم، برچسب x' به کمک رابطه (۳-۶) تعیین می‌گردد. فرض کنید *near(x)* نزدیک‌ترین همسایه به x در مجموعه داده آموزشی باشد. داده آزمایشی به کلاس "مثبت" تعلق می‌گیرد اگر:

$$score(x') = \frac{\sum_{i=1}^{K} d(x'.x_i^{NN})}{\sum_{i=1}^{K} d\left(d_i^{NN}.near(x_i^{NN})\right)} < \delta \qquad (۳-۶)$$

که $d(x'.x_i^{NN})$ فاصله داده آموزشی از i-اُمین داده موجود در مجموعه همسایگی‌اش و $d(d_i^{NN}.near(x_i^{NN}))$ فاصله i-اُمین داده موجود در مجموعه همسایگی داده آزمایشی تا نزدیک‌ترین همسایه به خودش در مجموعه داده آموزشی می‌باشد. نسبت گفته شده باید از آستانه

δ کمتر باشد تا داده آزمایشی برچسب "مثبت" دریافت کند (معمولاً در کاربردها از مقدار $\delta = 1$ استفاده می‌شود[۲۳]).

در این بخش نیز مشابه روش دسته‌بندی دو کلاسه، ما به هر داده آموزشی یک فیلد شمارنده اضافه می‌نماییم، یعنی $T = \{(x_i. c_i)\}_{i=1}^{N}$. برای این منظور ابتدا رابطه (۳-۶) را به صورت زیر تغییر می‌دهیم:

$$score(x') = \frac{\sum_{i=1}^{K} \frac{1}{d(x'. x_i^{NN})}}{\sum_{i=1}^{K} \frac{1}{d\left(d_i^{NN}. near(x_i^{NN})\right)}} > \frac{1}{\delta} \qquad (۳-۷)$$

که در این رابطه، صورت کسر بیانگر میزان شباهت داده آزمایشی به مجموعه همسایگی خود و مخرج کسر میزان نزدیکی نمونه‌های موجود در مجموعه همسایگی داده آزمایشی به سایر داده‌های آموزشی می‌باشد. سپس فیلد شمارنده نیز به عنوان ضریب به آن اضافه می‌کنیم:

$$score(x') = \frac{\sum_{i=1}^{K} \frac{c_i^{NN}}{d(x'. x_i^{NN})}}{\sum_{i=1}^{K} \frac{c_i^{NN(x_i^{NN})}}{d\left(d_i^{NN}. near(x_i^{NN})\right)}} > \frac{1}{\delta} \qquad (۳-۸)$$

که c_i^{NN} شمارنده متناظر با i-اُمین نمونه موجود در مجموعه همسایگی داده آزمایشی می‌باشد. روش به روز رسانی مقدار c_i در بخش بعدی خواهد آمد و همچنین مقدار δ را در فصل چهارم تعیین می‌نماییم.

۳-۵ تقویت مجموعه داده آموزشی

همانطور که گفته شد، ما یک مجموعه اولیه‌ای از نمونه‌های آموزشی تولید نمودیم اما برای افزایش دقت دسته‌بندی، نیاز به افزودن نمونه‌های بیشتر به مجموعه داده آموزشی داریم. از طرف دیگر، افزودن نمونه‌های بیشتر به معنی مصرف حافظه و محاسبات بیشتر در زمان دسته‌بندی می‌باشد. لذا به جای افزودن نمونه‌های بیشتر، ما فیلد جدیدی به نام شمارنده (با مقدار اولیه ۱) به هر نمونه آموزشی اضافه می‌نماییم. در زمان دسته‌بندی، اگر از الگوریتم دو کلاسه استفاده نماییم و داده آزمایشی به عنوان "مثبت" یا "منفی"، برچسب زده شود و یا اگر از الگوریتم تک کلاسه

استفاده کنیم و داده آزمایشی به دسته "مثبت" تعلق گیرد، آنگاه به ازای تمام نمونه‌های موجود در مجموعه همسایگی داده آزمایشی، چنانچه فاصله هر کدام از آن نمونه کمتر از آستانهٔ α بود، مقدار فیلد شمارندهٔ آن نمونه آموزشی به صورت زیر به روز رسانی می‌شود:

$$c_i = c_i + (1 - \frac{d(x'.x_i^{NN})}{\sum_{j=1}^{K} d(x'.x_j^{NN})}) \quad (۹-۳)$$

هرچه مقدار فیلد شمارندهٔ یک داده آموزشی بزرگتر باشد، به این معنا است که این داده به نمایندگی از تعداد بیشتری داده در مجموعه آموزشی حضور دارد و لذا از اهمیت بالاتری برخوردار است. در نتیجه، طبق روابط (۳-۵) و (۳-۸)، سهم بیشتری در تعیین برچسب یک داده جدید خواهد داشت. به همین ترتیب، نمونه‌های آموزشی با مقدار شمارنده کوچکتر، نقش کمتری را در تعیین برچسب داده جدید ایفا می‌نند. مقدار α به صورت تجربی برابر با ۰/۲ در نظر گرفته می‌شود. نتایج حاصل از افزوده شدن این فیلد و بهبود کارایی الگوریتم دسته‌بندی در فصل بعد نشان داده خواهد شد.

سیستم مطرح شده در این تحقیق قابلیت به کارگیری به صورت برخط را دارا می‌باشد. در سیستم برخط، هرگاه موردی به عنوان "غیر نرمال" تشخیص داده شود، منجر به نمایش کپچا می‌شوند و از طرف دیگر کاربر به آنها پاسخ نمی‌دهد (ربات‌ها)، به مجموعه آموزشی اضافه گردند. برای این منظور، لازم است ابتدا به دلیل محدودیت حافظه و سربار پردازشی در زمان دسته‌بندی، آستانه‌ای برای تعداد داده‌های آموزشی در نظر گرفته شود. سپس می‌توان آن داده را تحت دو شرط زیر به مجموعه آموزشی اضافه نمود:

۱) اگر اندازهٔ مجموعه آموزشی کمتر از محدودیت تعیین شده باشد، داده آزمایشی به مجموعه آموزشی افزوده می‌شود.

۲) در غیر این صورت، دو نمونه آموزشی از مجموعه نمونه‌های آموزشی که کمترین فاصله از یکدیگر دارند را یافته و انها را با هم ادغام می‌کنیم تا فضا برای نگه‌داری نمونه آزمایشی جدید باز شود. برای ادغام دو نمونه آزموزشی، میانگین بردار ویژگی‌های آن دو را محاسبه و فیلدهای شمارندهٔ آنها را با هم جمع می‌کنیم. یعنی دو نمونه $(x_i.c_i)$ و $(x_j.c_j)$ را از مجموعه داده آموزشی حذف و نمونه جدید $(\frac{x_i+x_j}{2}.c_i + c_j)$ و همچنین نمونه آزمایشی را به مجموعه آموزشی اضافه می‌نماییم.

فصل چهارم: ارزیابی

۴-۱ مقدمه

گرچه هر یک از روش‌های عنوان شده در فصول قبل با رویکرد متفاوتی به مقابله با کلیک‌های هرز می‌پردازند اما عدم وجود مجموعه داده عمومی از نشست‌های نرمال و غیر نرمال رفتار کاربران، موجود دشوار شدن عملیات ارزیابی و مقایسه این روش‌ها می‌شود. لذا عموماً ارزیابی‌ها به بررسی نمونه‌های غیر نرمال توسط افراد خبره و محاسبه معیارهایی نظیر دقت[1] محدود می‌شوند. ما نیز در این پژوهش با همین محدودیت مواجه بودیم و تنها امکان محاسبه این معیار را داشتیم، با این حال سعی نمودیم دقت روش‌های پیشنهادی را از جنبه‌های مختلف مورد ارزیابی و مقایسه قرار دهیم.

در فصل جاری، ابتدا به ارزیابی مجموعه داده آموزشی و مقایسه مجموعه داده اولیه و مجموعه داده آموزشی تقویت شده می‌پردازیم. سپس نحوه انتخاب پارامترها و کارایی دسته‌بندها را مورد ارزیابی قرار می‌دهیم و در نهایت، دقت روش‌های پیشنهادی را با یکی از آخرین و بهترین کارهای مرتبط مطرح شده مقایسه خواهیم نمود.

۴-۲ اعتبارسنجی متقابل K وجهی[2]

در بخش ۳-۳، روش استفاده شده جهت تولید داده‌های آموزشی را تشریح نمودیم. حال به ارزیابی مجموعه داده تولید شده از آن می‌پردازیم. برای این منظور، از روش اعتبارسنجی متقابل K وجهی استفاده می‌نماییم. در این روش، نمونه‌های آموزشی به صورت تصادفی به K بخش مساوی تقسیم می‌شوند. از یک بخش به عنوان داده ارزیابی و از (K-1) بخش دیگر به عنوان داده آموزشی استفاده می‌گردد. این فرآیند K بار تکرار می‌شود. بنابراین، از تمامی نمونه‌ها برای آموزش استفاده می‌شود و هر نمونه نیز یک بار برای ارزیابی مورد استفاده قرار می‌گیرد. در نهایت، مجموع یا میانگین نتایج هر دور به عنوان تخمین نهایی محاسبه می‌گردد. در این پژوهش، ما از مقدار K=10 یعنی اعتبارسنجی متقابل ۱۰ وجهی، به دلیل محبوبیت بیشتر در تحقیقات علمی استفاده کردیم.

[1] Precision
[2] K-Fold Cross-Validation

در ادامه، به ارزیابی مجموعه داده‌های تولید شده برای هر یک از روش‌های پیشنهادی (دو کلاسه و تک کلاسه) می‌پردازیم اما پیش از آن، مفاهیم به کار رفته در روش اعتبارسنجی متقابل را مرور می‌کنیم:

- **مثبت صحیح**[1]: این مقدار بیانگر تعداد نمونه‌هایی است که دسته واقعی آنها "مثبت" بوده و الگوریتم دسته‌بندی نیز دسته آنها را به درستی "مثبت" تشخیص داده است.
- **منفی کاذب**[2]: این مقدار بیانگر تعداد نمونه‌هایی است که دسته واقعی آنها "مثبت" بوده و الگوریتم دسته‌بندی، دسته آنها را به اشتباه "منفی" تشخیص داده است.
- **منفی صحیح**[3]: این مقدار بیانگر تعداد نمونه‌هایی است که دسته واقعی آنها "منفی" بوده و الگوریتم دسته‌بندی نیز دسته انها را به درستی "منفی" تشخیص داده است.
- **مثبت کاذب**[4]: این مقدار بیانگر تعداد نمونه‌هایی است که دسته واقعی آنها "منفی" بوده و الگوریتم دسته‌بندی، دسته آنها را به اشتباه "مثبت" تشخیص داده است.

با توجه به مفاهیم فوق، معیار "دقت دسته‌بندی"[5] به این صورت تعریف می‌شود:

$$CA = \frac{TP + TN}{TP + FN + TN + FP} \qquad (۱-۴)$$

معیار معرفی شده، مشهورترین و عمومی‌ترین معیار محاسبه کارایی الگوریتم‌های دسته‌بندی می‌باشد. این معیار، نشان دهندهٔ این حقیقت است که دسته‌بند طراحی شده قادر سات 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.

[1] True Positive (TP)
[2] False Negative (FN)
[3] True Negative (TN)
[4] False Positive (FP)
[5] Classification Accuracy

رابطه (۴-۱) در مسائل دسته‌بندی دو کلاسه مورد استفاده قرار می‌گیرد. در این مسائل، دو مقدار "مثبت صحیح" و "منفی صحیح" مهم‌ترین مقادیری هستند که باید بیشینه شوند. به طریق مشابه، در دسته‌بندی‌های تک کلاسه، رابطه (۴-۱) به صورت زیر تغییر می‌یابد:

$$CA = \frac{TP}{TP + FP} \qquad (۴-۲)$$

که در این وضعیت، برای کسب بیشترین دقت، تنها می‌بایست مقدار "مثبت صحیح" بیشینه شود.

برای تولید مجموعه آموزشی اولیه، ما از یک هفته لاگ جستجو و کلیک استفاده نمودیم (از تاریخ ۹۴/۰۹/۱۰ تا ۱۶/۰۹/۱۶) که شامل بیش از دو میلیون و سیصد هزار درخواست (پرس‌وجوها و کلیک‌ها) و بیشتر از نهصد و سی هزار نشست یکتا بود. پس از مدل‌سازی نشست‌های کاربران، مطابق روش گفته شده در بخش ۳-۳، مجموعه اولیه‌ای از داده‌های آموزشی تولید نمودیم. در دو بخش آتی، به ارزیابی این مجموعه برای دو دسته‌بند پیشنهاد شده (دو کلاسه و تک کلاسه) می‌پردازیم.

۴-۲-۱ دسته‌بند دو کلاسه

در روش دسته‌بندی دو کلاسه، ما مجموعه‌ای شامل ۱۰۰۰۰ نمونه آموزشی آماده نمودیم. برای تولید این مجموعه، همانطور که در بخش ۳-۳ تشریح شد، ابتدا یک مدل پنج بعدی از ویژگی‌های ۱) میانگین احتمال توالی فعالیت‌ها در زنجیره مارکوف، ۲) تعداد کل فعالیت‌های کاربر، ۳) نسبت تعداد پرس‌وجوها به کل فعالیت‌ها، ۴) نسبت تعداد کلیک‌های روی نتایج به کل فعالیت‌ها و ۵) نسبت تعداد کلیک‌های روی شماره صفحات به کل فعالیت‌ها، ایجاد می‌نماییم. سپس، فاصله ماهالانوبیس نشست‌ها را محاسبه نموده و ۵۰۰۰ مورد از نشست‌هایی که بیشترین فاصله را از میانگین داشته‌اند به عنوان نشست غیر نرمال و ۵۰۰۰ مورد از نشست‌هایی که کمترین فاصله را از میانگین داشته‌اند، به عنوان نشست نرمال در نظر می‌گیریم. در ادامه، هریک از این نشست‌ها مطابق بخش ۳-۲ مدل‌سازی شده و سپس نشست‌های غیر نرمال با برچسب "مثبت" یا

"یک" و نشست‌های نرمال با برچسب "منفی" یا "صفر" به مجموعه داده آموزشی اضافه می‌گردند.

پس از آماده‌سازی مجموعه آموزشی، الگوریتم اعتبارسنجی متقابل ۱۰ وجهی را به ازای مقادیر مختلف K (یا همان اندازه مجموعه همسایگی در الگوریتم KNN) بر روی داده‌های آموزشی تکرار نمودیم. در این حالت، فیلد شمارنده تمام داده‌های آموزشی دارای مقدار پیش‌فرض "۱" می‌باشد. نتایج حاصل شده در جدول ۴-۱ نشان داده شده است.

جدول ۴-۱: نتایج اعتبارسنجی متقابل ۱۰ وجهی روی مجموعه داده اولیه در دسته‌بندی دو کلاسه

	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
مثبت صحیح	۴۶۴۴	۴۶۰۲	۴۶۷۲	۴۷۱۸	۴۷۹۳	۴۸۰۲	۴۸۰۲
منفی کاذب	۱۹۶	۲۱۶	۱۸۲	۱۶۱	۱۱۱	۱۰۷	۱۰۶
منفی صحیح	۴۶۹۷	۴۶۸۱	۴۷۰۹	۴۷۶۷	۴۷۶۰	۴۷۵۹	۴۷۶۰
مثبت کاذب	۶۲	۶۹	۵۹	۴۸	۴۵	۴۴	۴۷
نامشخص	۴۰۱	۴۳۲	۳۷۸	۳۰۶	۲۹۱	۲۸۸	۲۸۵
دقت (٪)	۹۳/۴۱	۹۲/۸۳	۹۳/۸۱	۹۴/۸۵	۹۵/۵۳	۹۵/۶۱	۹۵/۶۲

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 فعالیت کاربران (از ۹۴/۰۹/۱۷ تا ۹۴/۰۹/۲۵) استفاده نموده و نشست‌های آن را مدل‌سازی می‌نماییم. سپس این مجموعه را به عنوان داده‌های آزمایشی به سیستم دسته‌بند دو کلاسه تزریق می‌کنیم تا در مورد دسته آنها تصمیم‌گیری شود. پس از دسته‌بندی تمامی نشست‌های متعلق به این مجموعه و به روز رسانی فیلد شمارنده نمونه‌های آموزشی، ما مجدداً مجموعه آموزشی تقویت شده را با کمک روش اعتبارسنجی متقابل ۱۰ وجهی مورد ارزیابی قرار دادیم. نتایج به دست آمده، در جدول ۴-۲ نشان داده شده است.

جدول ۴-۲: نتایج اعتبارسنجی متقابل ۱۰ وجهی روی مجموعه داده تقویت شده در دسته‌بندی دو کلاسه

	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
مثبت صحیح	۴۷۸۵	۴۷۸۸	۴۸۰۴	۴۸۲۸	۴۸۳۹	۴۸۴۸	۴۸۵۱
منفی کاذب	۱۰۲	۱۰۱	۹۵	۹۱	۹۰	۸۵	۸۳
منفی صحیح	۴۸۵۹	۴۸۶۱	۴۸۶۸	۴۸۷۴	۴۸۷۸	۴۸۸۰	۴۸۸۱
مثبت کاذب	۴۳	۴۳	۴۱	۳۸	۳۹	۳۷	۳۶
نامشخص	۲۱۱	۲۰۷	۱۹۲	۱۶۹	۱۵۴	۱۵۰	۱۴۹
دقت (٪)	۹۶/۴۲	۹۶/۴۹	۹۶/۷۲	۹۷/۰۲	۹۷/۱۷	۹۷/۲۸	۹۷/۳۲

مقایسه جدول ۴-۱ و ۴-۲ نشان می‌دهد که افزوده شدن فیلد شمارنده و انجام فرآیند تقویت داده‌های آموزشی، اولاً توانسته دقت دسته‌بندی را تا ۳٪ افزایش دهد و ثانیاً ما را از انتخاب مقدار مناسب برای پارامتر K (اندازه مجموعه همسایگی) که یکی از چالش‌های روش KNN می‌باشد، بی‌نیاز سازد. زیرا همانطور که در سطر آخر جدول ۴-۲ مشاهده می‌شود دقت محاسبه شده به ازای مقدار مختلف اندازه مجموعه همسایگی در مقیسه با جدول ۴-۱، نوسانات بسیار کمی دارد. بنابراین، ما صرفاً بر اساس نتایج جدول ۴-۱، از مقدار K=5 استفاده می‌کنیم، زیرا دقت روش تا پیش از آن افزایش می‌یابد اما پس از آن با افزایش K، تقریباً ثابت باقی می‌ماند. با این حساب، دقت روش پیشنهادی در حالت دسته‌بندی دو کلاسه برابر با ۹۵/۶۵٪ شده است.

۴-۲-۲ دسته‌بندی تک کلاسه

در روش دسته‌بندی تک کلاسه نیز از همان مجموعه داده آموزشی تولید شده برای حالت دو کلاسه استفاده می‌نماییم، با این تفاوت که در اینجا تنها از نمونه‌های مثبت استفاده می‌کنیم. بنابراین با مجموعه داده اولیه‌ای شامل ۵۰۰۰ نمونه مثبت که همگی دارای فیلد شمارنده‌ای با مقدار "۱" هستند، ارزیابی را انجام می‌دهیم. نتایج حاصل از روش اعتبارسنجی ۱۰ وجهی برای

مقادیر مختلف اندازه همسایگی K در جدول ۴-۳ نمایش داده شده است.

جدول ۴-۳: نتایج اعتبارسنجی متقابل ۱۰ وجهی روی مجموعه داده اولیه در دسته‌بندی تک کلاسه

	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 روزرسانی فیلد شمارندهٔ آنها مورد بررسی قرار می‌دهیم. برای این منظور، از لاگ یک هفته‌ای تراکنش کاربران (از تاریخ ۹۴/۰۹/۱۷ تا ۹۴/۰۹/۲۵) استفاده نموده و آنها را مدل‌سازی کردیم. سپس این مجموعه را به عنوان داده‌های آزمایشیبه سیستم دسته‌بند تک کلاسه تزریق نمودیم. پس از دسته‌بندی داده‌های این مجموعه و به روز رسانی فیلد شمارنده نمونه‌های آموزشی، ما مجدداً مجموعه آموزشی تقویت شده را با کمک روش اعتبارسنجی متقابل ۱۰ وجهی مورد ارزیابی قرار دادیم. نتایج حاصل، در جدول ۴-۴ نمایش داده شده است.

جدول ۴-۴: 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 ۴-۳ و ۴-۴ نشان می‌دهد که ما در دسته‌بند تک کلاسه نیز توانستیم دقت دسته‌بند پیشنهادی را به میزان تقریباً ۲٪ بهبود دهیم. همچنین مشابه با آنچه در مورد

دسته‌بند دو کلاسه گفته شد، دقت دسته‌بند در وضعیت دوم، یعنی پس از تقویت داده‌های آموزشی، به ازای مقادیر مختلف اندازه همسایگی در حدود ۹۷٪ ثابت باقی می‌ماند، لذا با توجه به وضعیت نخست، ما مقدار K=5 را به عنوان اندازه مجموعه همسایگی در نظر می‌گیریم. بنابراین، دقت روش تک کلاسهٔ پیشنهادی برابر با ۹۶/۵۴٪ می‌شود.

۳-۴ کارایی الگوریتم دسته‌بندی

در این بخش، کارایی الگوریتم‌های پیشنهادی با لاگ یک هفته (از ۹۴/۰۹/۱۷ تا ۹۴/۰۹/۲۵) که شامل بیش از ۲ میلیون رکورد بود، ارزیابی می‌شود. برای این منظور، ابتدا نشست‌های کاربران را مدلسازی کرده و سپس آنها را به الگوریتم‌های دسته‌بندی که در بخش ۳-۴ به معرفی آنها پرداختیم، تزریق می‌نماییم. پس از انجام عملیات دسته‌بندی، هریک از نشست‌ها یک امتیاز دریافت می‌کنند. ما در گام نخست، فرکانس امتیازهای تخصیص داده شده به هر نشست را در دو حالت قبل و بعد از اعمال تغییرات پیشنهادی، در بازه‌های مختلف به دست آورده و نمودار آنها را ترسیم می‌نماییم. آنگاه از نتایج مشاهده شده برای انتخاب پارامترهای حد آستانه در روابط (۳-۵) و (۳-۸) استفاده می‌کنیم. در گام بعد، تعدادی از نشست‌های موجود در هر محدوده امتیاز را مورد بررسی بیشتر قرار داده و دقت دسته‌بندی را ارزیابی می‌کنیم.

بنا بر آنچه گفته شد، در ادامه ابتدا به بررسی کارایی دسته‌بند دو کلاسه و سپس دسته‌بند تک کلاسه خواهیم پرداخت.

۳-۴-۱ دسته‌بند دو کلاسه

در این قسمت، ابتدا به مقایسه الگوریتم پایه‌ای که برای دسته‌بندی دو کلاسه استفاده شد (رابطهٔ ۳-۴) با روش پیشنهاد شده (رابطهٔ ۳-۵) می‌پردازیم. این الگوریتم‌ها در بخش ۳-۴-۱ تشریح شدند که طبق آنها، امتیازهای تخصیص یافته به هر نشست پس از دسته‌بندی در بازهٔ [۰-۱] قرار می‌گیرد. بنابراین، ما این بازه را به ده بخش تقسیم کرده و فرکانس امتیازهای تخصیص داده شده به هر نشست را در هریک از بازه‌ها محاسبه می‌نماییم. نتایج به دست آمده در شکل ۴-۱ نشان داده شده است.

شکل ۴-۱: توزیع فرکانس امتیازهای حاصل از دسته‌بندی در بازه‌های مختلف برای دسته‌بندی دو کلاسه

پس از رسم توزیع امتیازهای به دست آمده از دسته‌بندی نشست‌ها (شکل ۴-۱) مشاهده می‌کنیم که در روش پایه، ۹۲٪ از نشست‌ها، امتیازی کمتر از ۰/۱ به دست آوردند و ۷٪ از نشست‌ها امتیازی بیشتر از ۰/۹ کسب نمودند. در حالی که در روش پیشنهادی، ۹۵٪ نشست‌ها امتیازی کمتر از ۰/۱ و ۴/۵٪ از نشست‌ها، امتیازی بالاتر از ۰/۹ به دست آوردند. با توجه به نحوه کارکرد الگوریتم، می‌دانیم هر چه امتیاز محاسبه شده برای یک نشست بیشتر باشد (به یک نزدیک باشد)، احتمال غیر نرمال بودن آن افزایش می‌یابد و از طرف دیگر، هرچه این امتیاز به صفر نزدیک باشد، نشست مربوطه نرمال خواهد بود. بنابراین، ما مقدار ۰/۱ را برای حد آستانه انسانی و مقدار ۰/۹ را برای حد آستانه رباتی در الگوریتم دسته‌بندی دو کلاسه انتخاب می‌کنیم.

بنا بر آنچه گفته شد، در روش پیشنهادی، تعداد نشست‌های غیر نرمال ۲/۵٪ کمتر شده است. لذا این سوال مطرح می‌شود که آیا روش پیشنهادی از دقت کمتری در تشخیص نشست‌های غیر نرمال برخوردار است؟ برای پاسخ به این سوال، ما دست به آزمایشات بیشتری زدیم. برای این منظور، ابتدا به ازای هر یک از روش‌ها، به صورت تصادفی ۲۰۰ نشست که امتیاز آنها در بازه [۱-۰/۹] بود، انتخاب نمودیم. سپ به صورت دستی، صحت تشخیص آنها را با در نظر گرفتن فاکتورهایی نظیر تعداد کلیک‌های تکرای و ارتباط پرس‌وجوهای ارسالی با لینک‌های کلیک‌شده بررسی نمودیم. نتیاج این بررسی، در جدول ۴-۵ نشان داده شده است.

جدول ۴-۵: دقت الگوریتم دو کلاسه در بازه‌های امتیازی مختلف

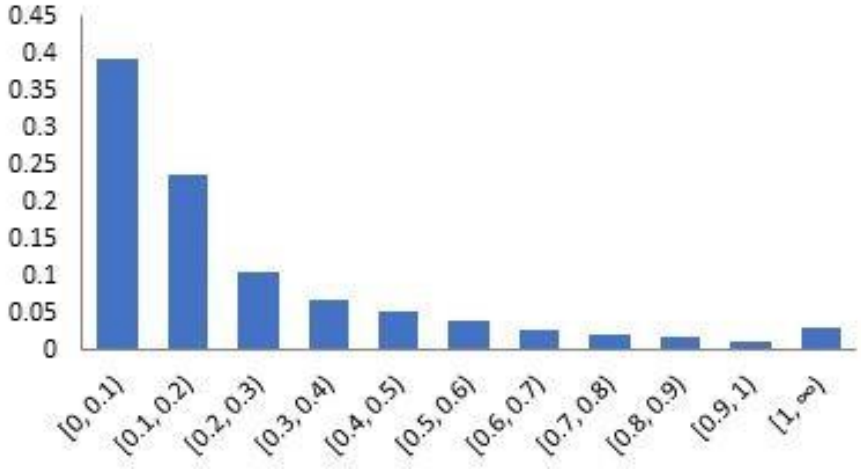
	تعداد نشست‌ها	نشست‌های درست تشخیص داده شده	دقت (٪)
روش پایه	۲۰۰	۱۶۲	۸۱
روش پیشنهادی	۲۰۰	۱۸۳	۹۱/۵

نتایج حاصل از بررسی دستی رفتار نشست‌ها نشان می‌دهد که الگوریتم پیشنهادی می‌تواند با دقت ۹۱/۵٪ نشست‌های غیر نرمال را شناسائی نماید که در مقایسه با الگوریتم پایه ۱۰/۵٪ بهبود از خود نشان می‌هد. بنابراین می‌توان نتیجه گرفت که افزودن پارامتر شمارنده، توانسته دقت روش دو کلاسه را بیش از ۱۰٪ افزایش دهد. و در پاسخ به سؤال مطرح شده باید گفت، روش پایه به این دلیل که مثبت کاذب بالاتری داشته، منجر به موارد غیر نرمال بیشتری شده است در حالی که روش بهبود یافته با دقت بالاتری نشست‌های غیر نرمال را شناسائی می‌نماید. بررسی نشست‌هایی که به اشتباه به عنوان غیر نرمال، تشخیص داده شدند نشان داد که نمونه‌های نرمال در مجموعه آموزشی نمی‌توانند به خوبی تمام جنبه‌های رفتاری کاربران نرمال را پوشش دهند، زیرا کاربران نرمال از تنوع رفتاری بالایی برخوردار هستند. این مسئله و همچنین این حقیقت که هدف ما تشخیص نشست‌های غیر نرمال است تا نرمال، موجب شد تا ما تصمیم بگیریم که با به کارگیری یک الگوریتم دسته‌بند تک کلاسه، تنها از نمونه‌های مثبت استفاده نموده و با دقت بالاتری موارد غیر نرمال را تشخیص دهیم. به این ترتیب، روش مطرح شده در بخش ۳-۴-۲ را ارائه نمودیم. در بخش بعد، به بررسی کارایی الگوریتم پیشنهادی برای دسته‌بند تک کلاسه می‌پردازیم.

۴-۳-۲ دسته‌بند تک کلاسه

در این بخش، به ارزیابی روش تک کلاسهٔ پیشنهادی می‌پردازیم. با توجه به الگوریتم مطرح شده در دسته‌بند تک کلاسه و به طور دقیق‌تر از رابطه (۳-۸)، می‌دانیم که امتیاز محاسبه شده برای هر نشست می‌تواند در بازهٔ $[0 - \infty)$ متغیر باشد. در اینجا نیز مشابه قبل، ابتدا فرکانس

امتیاز تخصیص یافته به هر نشست را در بازه‌های مختلف به دست می‌آوریم. در روش تک کلاسه نیز ما طول همهٔ بازه‌ها را ۰/۱ در نظر می‌گیریم ولی جهت محدود کردن تعداد بازه‌ها، امتیازات بزرگتر یا مساوی با ۱ را به صورت مجتمع در یک بازه در نظر می‌گیریم. نتایج حاصل در شکل ۴-۲ نشان داده شده است.



شکل ۴-۲: توزیع فرکانس امتیازهای حاصل از دسته‌بندی در بازه‌های مختلف برای دسته‌بند تک کلاسه

پس از تقسیم‌بندی نشست‌های کاربران، در گام دوم به آنالیز بیشتر نشست‌ها می‌پردازیم. در روش تک کلاسه، هدف ما این است که نمونه‌های مثبت (موارد غیر نرمال) را به درستی و با دقت بیشتری تشخیص دهیم. از طرف دیگر، بنا بر الگوریتم پیشنهاد شده می‌دانیم که هرچه امتیاز نهایی یک نشست بیشتر باشد، به احتمال زیاد آن نشست، رفتاری غیر نرمال و رباتی داشته است. از اینرو، برای انجام آزمایشات دقیق‌تر، 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 ۴-۶ نمایش داده شده است.

جدول ۴-۶: دقت الگوریتم تک کلاسه در بازه‌های امتیازی مختلف

محدوده امتیاز	تعداد نشست‌ها	تعداد غیرنرمال‌ها	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
مجموع	۲۰۰	۱۲۴	۶۲

با توجه به نتایج ارائه شده در جدول ۴-۶ و اینکه شناسائی موارد هرز نیاز به دقت بالایی دارد، لذا ما نشست‌هایی که امتیاز بالاتر از ۰/۸ داشته‌اند را به عنوان نشست غیر نرمال در نظر گرفتیم و بنابراین مقدار ۱/۲۵ = ۰/۸ / ۱ را برای پارامتر δ در رابطه (۳-۸) لحاظ نمودیم. با این حساب، دقت روش برابر با ۹۶/۶۱٪ (که از نسبت ۱۱۴ نشست غیر نرمال درست تشخیص داده شده به ۱۱۸ نشست مورد بررسی به دست می‌آید) می‌شود. در نتیجه کلیه کلیک‌های انجام شده در این نشست‌ها به عنوان کلیک هرز قلمداد می‌شوند.

در دو بخش اخیر، به ارزیابی مجموعه داده آموزشی و همچنین کارایی الگوریتم‌های دسته‌بندی دو کلاسه و تک کلاسه پرداختیم. ارزیابی‌ها نشان داده که برای مسئله تعریف شده در این پژوهش، روش دسته‌بندی تک کلاسه به دقت بالاتری منتج می‌شود اما این نتیجه الزاماً در مورد تمامی مسائل دسته‌بندی صادق نیست. در مسئله مورد تحقیق ما، به دلیل تنوع بسیار بالای رفتار کاربران نرمال، نمونه‌های منفی به خوبی نمی‌توانند تمام جنبه‌های یک رفتار نرمال را پوشش دهند، لذا مطرح نمودن الگوریتم دسته‌بندی تک کلاسه، ضمن اینکه ما را از نگه‌داری نمونه‌های نرمال بی‌نیاز می‌سازد، در صورت به کارگیری سیستم در حالت برخط، سربار به روز رسانی نمونه‌های منفی را نیز ندارد (زیرا اکثر نشست‌ها نرمال هستند و به روز رسانی مجموعه آموزشی به

ازای هر فعالیت از یک نشست نرمال به عنوان یک سربار برای سیستم در نظر گرفته می‌شود). از اینرو، در مسئله ما الگوریتم تک کلاسه بهره‌وری بالاتری داشته و مناسب‌تر می‌باشد.

۴-۴ مقایسه با کارهای قبلی

در نهایت، کارایی دو الگوریتم پیشنهادی را با روش گراف دو بخشی مطرح شده در [۱۹] جهت شناسائی کلیک‌های هرز مقایسه می‌کنیم. بر اساس روش مذکور که به طور مفصل در بخش ۲-۳-۳ به معرفی آن پرداختیم، ابتدا نشست‌های کاربران را به صورت توالی‌های سه‌تایی از نوع فعالیت، هدف فعالیت و اختلاف زمانی آن را با فعالیت قبلی‌اش مدل کرده و سپس الگوریتم گراف دو بخشی کاربر-نشست را برای شناسائی تعداد بیشتری از نشست‌های مشکوک اعمال می‌کنیم. در انتها ما نشست‌هایی را که امتیازی بالاتر از ۰/۹ داشته‌اند به عنوان نشست غیر نرمال در نظر می‌گیریم. آنگاه به صورت تصادفی ۲۰۰ نمونه از آنها را انتخاب و درستی تشخیص آنها را به صورت دستی بررسی نمودیم که دقتی برابر با ۹۳/۵٪ (تشخیص ۱۸۷ نمونه از ۲۰۰ نمونه) حاصل شد. مقایسه دقت محاسبه شده با نتایج ارائه شده در جدول ۴-۵ و ۴-۶، نشان می‌دهد که روش دو کلاسه پیشنهاد شده در این پژوهش در حدود ۲ درصد و روش تکلاسه در حدود ۳/۱ درصد عملکرد بهتری روی داده‌های ما داشته‌اند. البته لازم است به این نکته نیز توجه نمود که گراف دو بخشی بیشتر یک روش برون‌خط محسوب می‌شود، در حالی که روش پیشنهادی در این پایان‌نامه قادر است تا به صورت برخط، با مصرف حافظه پایین و انجام محاسباتی بسیار ساده‌تر به شناسائی کلیک‌های هرز بپردازد.

فصل پنجم: نتیجه‌گیری

۵-۱ مقدمه

در دهۀ اخیر مسئله کلیک‌های هرز به عنوان یک چالش اساسی در شبکه‌های تبلیغاتی، موتورهای جستجو و اساساً هر سیستم آنلاینی که از کلیک‌ها به عنوان بازخوردی از رفتار کاربران استفاده می‌نماید، مطرح شده است. لذا در این پایان‌نامه، ما روش جدیدی جهت تشخیص کلیک‌های هرز با رویکرد شناسائی نشست‌های غیر نرمال کاربران ارائه نمودیم. سپس روش پیشنهادی را در یک موتور جستجوی بومی مورد بررسی و تحلیل قرار دادیم زیرا این مسئله در موتورهای جستجو به دلیل تأثیر نامطلوب در رتبه‌بندی نتایج جستجو و افزایش زمان پاسخگویی به کاربران حقیقی از اهمیت بسیار بیشتری برخوردار می‌باشد.

روش‌هایی که تاکنون در این حوزه مطرح شده‌اند، هر یک به جنبه خاصی از رفتارهای غیرنرمال پرداخته و تنها قادر به شناسائی حملات انجام شده در آن دسته هستند اما ما در این پژوهش، سعی نمودیم جنبه‌های مختلفی از رفتار ترافیک‌های غیرنرمال را به صورت مجموعه‌ای از ویژگی‌ها در سه سطح نشست، کاربر و آدرس IP با یکدیگر ترکیب نموده و با کمک تکنیک‌های دسته‌بندی پیشنهاد شده، به شناسائی کلیک‌های هرز بپردازیم. همچنین بیشتر روش‌های قبلی به صورت برون‌خط کار می‌کردند اما سیستم مطرح شده در این پایان‌نامه، قابلیت به کارگیری برخط را نیز دارا می‌باشد گرچه این قابلیت هنوز عملیاتی نشده است.

۵-۲ یافته‌های تحقیق

روش‌هایی که تا کنون جهت شناسائی کلیک‌های هرز در موتورهای جستجو مطرح شده‌اند، معمولاً در یکی از سه گروه ۱) شناسائی ترافیک‌های غیر نرمال، ۲) شناسائی کاربران غیر نرمال و ۳) شناسائی نشست‌های غیر نرمال قرار می‌گیرند. روش پیشنهاد شده در این پایان‌نامه در دستۀ سوم جای می‌گیرد. بنابراین عمده ویژگی‌های معرفی شده، از نشست جاری کاربر محاسبه می‌شود اما با این وجود، ما سعی نمودیم با افزودن ویژگی‌هایی در سطح رفتاری کاربران (مجموع نشست‌های کاربر) و همچنین رفتار آدرس‌هایIP متناظر با آن نشست، دامنۀ تشخیص روش پیشنهادی را افزایش دهیم.

پس از مدل‌سازی نشست‌های کاربران به کمک ویژگی‌های مطرح شده، الگوریتم‌های پیشنهادی خود را که مبتنی بر الگوریتم KNN می‌باشد، مطرح نمودیم. در روش‌های پیشنهادی، ما توانستیم با افزودن یک پارامتر سادهٔ "شمارنده" و محدود نمودن تعداد داده‌های آموزشی بر مشکلات الگوریتم اولیه KNN که شامل حافظهٔ مصرفی و حجم زیاد محاسبات بود، فائق آییم. اما از طرف دیگر با ایجاد مکانیزم به روز رسانی مجموعه آموزشی، سعی نمودیم همواره امکان افزودن نمونه‌های جدید به سیستم وجود داشته باشد.

در بخش روش‌های پیشنهاد شده، ما ابتدا یک الگوریتم دسته‌بندی دو کلاسه را پیشنهاد دادیم. این الگوریتم که از مجموعه‌ای حاوی نمونه‌هایی از هر دو دستهٔ مثبت و منفی به عنوان داده‌های آموزشی تغذیه می‌کند، توانست کلیک‌های هرز را با دقت ۹۵/۵٪ شناسائی نماید. گرچه روش دو کلاسه، می‌تواند با دقت خوبی کار کند اما ما یک گام پیشتر رفته و با نگه‌داری تنها نمونه‌های مثبت، استفاده از یک الگوریتم تک کلاسه را پیشنهاد دادیم. این الگوریتم نسبت به روش دو کلاسه، تنها از تنها نمونه‌های مثبت در مجموعهٔ آموزشی خود بهره میگیرد لذا ما می‌توانیم با کاهش تعداد نمونه‌های آموزشی و به تبع آن کاهش حجم محاسبات به کارآمدی و دقتی بیش از پیش برسیم. ارزیابی‌ها نشان می‌دهد که روش تک کلاسه می‌تواند با دقت ۹۶/۶۱ درصدی، کلیک‌های هرز را شناسائی نماید که این میزان نسبت به الگوریتم دو کلاسه حدود ۱/۱ درصد بهبود از خود نشان می‌دهد. همچنین موجب شد که ما نسبت به کارهای پیشین به بهبود دقت تشخیص ۳/۱ درصدی دست یابیم.

بنابراین می‌توان یافته‌های اصلی این پژوهش را در موارد زیر خلاصه نمود:

- شناسائی محدودهٔ متنوعی از حملات کلیک‌های هرز با به کارگیری و ترکیب ویژگی‌های مختلف رفتاری کاربران نرمال و ربات‌ها
- شناسائی کلیک‌های هرز با دقت قابل قبول و بیش از روش‌های قبلی
- امکان به کارگیری روش‌های پیشنهادی به صورت برخط (در مقایسه با روش‌های قبلی که همگی به صورت برون‌خط عمل می‌کنند)

۵-۳ پشنهادها

در ادامهٔ این پژوهش می‌توان موارد زیر را به عنوان کارهای آتی در این حوزه پیشنهاد داد:

- عملیاتی‌سازی روش‌های پیشنهادی به صورت برخط و بررسی عملکرد آن
- بررسی و شناسائی حملات مختلف بر اساس نشست‌های هرز تشخیص داده شده
- افزودن ویژگی‌های مرتبط با لینک‌های کلیک شده (نظیر ساختار پیوندی و محتوایی صفحهٔ مذکور)

واژه نامه فارسی به انگلیسی

اعتبارسنجی متقابل کا وجهی	*K-Fold Cross-Validation*
برخط	*Online*
برون‌خط	*Offline*
پرداخت به ازای هر کلیک	*Pay Per Click*
پرس‌وجوی هرز	*Query Spam*
تأثیر	*Impression*
تبلیغ‌کننده	*Advertiser*
دقت	*Precision*
دقت دسته‌بندی	*Classification Accuracy*
ربات	*Bot*
سرور نام میزبان	*Domain Name Server*
شبکه تبلیغ	*Ad Network*
کلیک کلاهبردارانه	*Click Fraud*
ماهالانوبیس	*Mahalanobis*
مثبت صحیح	*True Positive*
مثبت کاذب	*False Positive*
مجموعه آموزشی	*Training Set*
معامله به سود	*Arbitrage*
منتشرکننده	*Publisher*
منفی صحیح	*True Negative*
منفی کاذب	*False Negative*
نمونه آزمایشی	*Test Point*
نمونه آموزشی	*Training Point*

واژه نامه انگلیسی به فارسی

شبکه تبلیغ	*Ad Network*
تبلیغ‌کننده	*Advertiser*
معامله به سود	*Arbitrage*
ربات	*Bot*
دقت دسته‌بندی	*Classification Accuracy*
کلیک کلاهبردارانه	*Click Fraud*
سرور نام میزبان	*Domain Name Server*
منفی کاذب	*False Negative*
مثبت کاذب	*False Positive*
تأثیر	*Impression*
اعتبارسنجی متقابل کا وجهی	*K-Fold Cross-Validation*
ماهالانوبیس	*Mahalanobis*
برون خط	*Offline*
برخط	*Online*
پرداخت به ازای هر کلیک	*Pay Per Click*
دقت	*Precision*
منتشرکننده	*Publisher*
پرس‌وجوی هرز	*Query Spam*
نمونه آزمایشی	*Test Point*
نمونه آموزشی	*Training Point*
مجموع آموزشی	*Training Set*
منفی صحیح	*True Negative*
مثبت صحیح	*True Positive*

فهرست مراجع

[1] M. Marchiori, "The Quest for Correct Information on the Web: Hyper Search Engines.," *Comput. Networks*, vol. 29, no. 8–13, pp. 1225–1236, 1997.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *Stanford Digital Library Technologies Project*, 1998.

[3] "I. A. Board. 2013 Internet Advertising Revenue Report." [Online]. Available: http://www.iab.net/.

[4] D. Szetela and J. Kerschbaum, *Pay-Per-Click Search Engine Marketing: An Hour a Day*. Alameda, CA, USA: SYBEX Inc., 2010.

[5] N. Daswani and M. Stoppelman, "The Anatomy of Clickbot.A," in *Proceedings of the 1st Conference on First Workshop on Hot Topics in Understanding Botnets*, p. 11, 2007.

[6] B. Miller, P. Pearce, C. Grier, C. Kreibich, and V. Paxson, "What's Clicking What? Techniques and Innovations of Today's Clickbots," in *Proceedings of the 8th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pp. 164–183, 2011.

[7] S. A. Alrwais, A. Gerber, C. W. Dunn, O. Spatscheck, M. Gupta, and E. Osterweil, "Dissecting Ghost Clicks: Ad Fraud via Misdirected Human Clicks," in *Proceedings of the 28th Annual Computer Security Applications Conference*, pp. 21–30, 2012.

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

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

[10] V. Dave, S. Guha, and Y. Zhang, "Measuring and Fingerprinting Click-spam in Ad Networks," in *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pp. 175–186, 2012.

[11] V. Dave, S. Guha, and Y. Zhang, "ViceROI: Catching Click-spam in Search Ad Networks," in *Proceedings of the ACM SIGSAC Conference on Computer & Communications Security*, pp. 765–776, 2013.

[12] A. Juels, S. Stamm, and M. Jakobsson, "Combating Click Fraud via Premium Clicks," in *Proceedings of 16th USENIX Security Symposium on USENIX*

Security Symposium, pp. 2:1–2:10, 2007.

[13] A. Metwally, D. Agrawal, A. El Abbad, and Q. Zheng, "On Hit Inflation Techniques and Detection in Streams of Web Advertising Networks," in *Proceedings of the 27th International Conference on Distributed Computing Systems*, p. 52, 2007.

[14] O. Stitelman, C. Perlich, B. Dalessandro, R. Hook, T. Raeder, and F. Provost, "Using Co-visitation Networks for Detecting Large Scale Online Display Advertising Exchange Fraud," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,*, pp. 1240–1248, 2013.

[15] F. Soldo and A. Metwally, "Traffic anomaly detection based on the IP size distribution," in *Proceedings of the IEEE INFOCOM*, pp. 2005–2013, 2012.

[16] 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.

[17] F. Yu, Y. Xie, and Q. Ke, "SBotMiner: Large Scale Search Bot Detection," in *Proceedings of the 3th ACM International Conference on Web Search and Data Mining*, pp. 421–430, 2010.

[18] 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.

[19] X. Li, M. Zhang, Y. Liu, S. Ma, Y. Jin, and L. Ru, “Search Engine Click Spam Detection based on Bipartite Graph Propagation,” in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pp. 93–102, 2014.

[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] R. A. Costa, R. J. G. B. de Queiroz, and E. R. Cavalcanti, “A Proposal to Prevent Click-Fraud Using Clickable CAPTCHAs,” in *Proceedings of the 2012 IEEE 6th International Conference on Software Security and Reliability Companion*, pp. 62–67, 2012.

[22] T. Cover and P. Hart, “Nearest Neighbor Pattern Classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Sep. 2006.

[23] W. Yousef, M. , Najami, N. and Khalifav, “A comparison study between one-

class and two-class machine learning for MicroRNA target detection," *Journal of Biomedical Science and Engineering*, pp. 247–252, 2010.

[24] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond Accuracy, F-score and ROC: A Family of Discriminant Measures for Performance Evaluation," in *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*, pp. 1015–1021, 2006.

Abstract

Most of today's internet services utilize user feedback (clicks) to improve the quality of their services. For example, search engines use click information as a key factor in document ranking. As a result, some websits cheat to get a higher rank by fraudulently increasing clicks to their pages. Since these clicks are not performed by real users, this phenomenon is known "click spam". Cheating websites can generate fake clicks by hiring peoples or using "bots". Bots are automated software programs issuing several queries or producing excessive clicks. So, the problem of distinguishing bot-generated traffic from the user traffic is critical for search engines. In the research, we propose novel classification-based techniques to identify fraudulent clicks. The proposed algorithms provide both effectiveness and efficiency for detecting click spam in a practical manner. We first model user sessions as a set of numerical features. These features are selected such that they can cover many aspect of normal and abnormal behavior and distinguish bot/human from each other as well as possible. Then, we describe our classification techniques which include a two-class and a one-class classification algorithm. After applying these approachs, we can detect normal/abnormal user sessions and as a result, click spams. Finally, we analyze our methods with real logs of a Persian search engine. Experimental results show that the proposed algorithms can detect fradudelent clicks with a precision of up to 96% which outperform the previous work by 3.1%.

key words *:Click spam, botnet detection, machine learning, k-Nearest Neighbour*

Yazd University

Faculty of Electrical and Computer Engineering

A Thesis Submitted in Partial Fulfillment of the Requirements for the Master Degree in Computer Engineering

Title

Click spam detection in Persian web space

Supervisor

Dr. Sajjad Zarifzadeh

Advisor

Dr. Vali Derhami

By

Mahdieh Fallah

February 2016