ارائهٔ راه‌کار برای مقابله با فریب ایجادشده به‌وسیلهٔ ربات‌ها به‌منظور بهبود رتبه‌بندی ترافیکی تارنماها

زهرا عبدی[۱]، مجتبی مازوچی*[۲] و محمدعلی پورمینا[۳]

[۱و۳]دانشکده برق و کامپیوتر، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران

[۲]پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

چکیده

با گسترش اینترنت و فضای وب، برقراری ارتباط و کسب اطلاعات در بین افراد از شکل سنتی و اولیهٔ خود فاصله گرفته و به درون تارنماها کشیده شده است. همچنین فضای جهانی وب، فرصت بزرگی را برای کسب و کارها فراهم می‌کند تا ارتباط خود را با مشتری بهبود ببخشند و بازار خود را در دنیای برخط گسترش دهند. کسب و کارها برای بررسی میزان بازدید و محبوبیت سایت‌هایشان از معیاری به نام رتبه‌بندی ترافیکی استفاده می‌کنند. رتبه‌بندی ترافیکی میزان بازدیدکنندگان یک سایت را اندازه گرفته و براساس همین آمار، رتبه‌ای را به سایت اختصاص می‌دهد. یکی از مهم‌ترین چالش‌های موجود در رتبه‌بندی، ایجاد ترافیک جعلی تولیدشده به‌وسیلهٔ برنامه‌های کاربردی به نام ربات است. ربات‌ها اجزای نرم‌افزاری مخرب مورد استفاده برای تولید هرزنامه‌ها، راه‌اندازی حملات مختل‌کنندهٔ سامانه، فیشینگ، سرقت هویت و خروج اطلاعات و دیگر فعالیت‌های غیر قانونی هستند تاکنون روش‌های مختلفی برای شناسایی و کشف ربات صورت گرفته است. در این پژوهش، شناسایی ربات‌ها از طریق تحلیل و پردازش لاگ دسترسی وب سرور و استفاده از روش‌های داده‌کاوی، انجام می‌شود. نتایج تجربی نشان می‌دهد که روش پیشنهادی در این پژوهش با کشف ویژگی‌های جدید و معرفی شرط جدید در برچسب‌گذاری نشست‌ها، باعث بهبود دقت در شناسایی ربات‌ها و در نتیجه ایجاد بهبود در رتبه‌بندی ترافیکی تارنماها نسبت به کارهای پیشین شده است.

واژگان کلیدی: رتبه‌بندی ترافیکی، شناسایی ربات، برچسب‌گذاری نشست، لاگ دسترسی وب سرور، داده‌کاوی

Representing a method to identify and contrast with the fraud which is created by robots for developing websites' traffic ranking

Zahra Abdi[1], Mojtaba Mazoochi*[2] & MohammadAli Pourmina[3]

[1,3]Faculty of Electrical and Computer Engineering, Islamic Azad University, Science and Research Branch, Tehran, Iran

[2]Assistant Professor, ICT Research Institute, Tehran, Iran

Abstract

With the expansion of the Internet and the Web, communication and information gathering between individual has distracted from its traditional form and into web sites. The World Wide Web also offers a great opportunity for businesses to improve their relationship with the client and expand their marketplace in online world. Businesses use a criterion called traffic ranking to determine their site's popularity and visibility. Traffic ranking measures the amount of visitors to a site and based on these statistics, allocates a ranking to the site. One of the most important challenges in the ranking is the creation of fake traffic that generated by applications called robots. Robots are malicious software components that used to generate spam, set up distributed denial of services attacks, fishing, identity theft, removal of information and other illegal activities. there are already several ways to identify and discover the robot. According to Doran et al., The identification methods are divided into two categories: offline and real-time. The offline detection method is divided into three categories:

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

● تاریخ ارسال مقاله: ۱۳۹۸/۰۴/۰۵ ● تاریخ پذیرش: ۱۳۹۸/۱۰/۲۱ ● تاریخ انتشار: ۱۴۰۰/۲۰/۲۹ ● نوع مطالعه: پژوهشی

Syntactical Log Analysis, Traffic Pattern Analysis, and Analytical Learning Techniques. The real-time method is performed by the Turing test system. In this research, the identification of robots is done through the offline method by analysis and processing of access logs to the web server and the use of data mining techniques. In this method, first, the features of each session are extracted, then generally these sessions are labeled with three conditions into two categories of human and robot. Finally, by using data mining tool, web robots are detected. In all previous studies, the features are extracted from each sessions, for example in first studies, Tan&Kumar extracted 25 features of sessions. After that Bomhardt et al. used 34 features to identify the robots. In 2009 Stassopoulou et al. used 6 features that was extracted from sessions and so on. But in this research, features are extracted from sessions of a unique user. Experimental results show that the proposed method in this research, by discovering new features and introducing a new condition in session labeling, improves the accuracy of identifying robots and moreover, improves the ranking of web traffic from previous work.

Keywords: Traffic Ranking, Robot Detection, Session Labeling, Web Server Access Log, Data Mining

۱- مقدمه

مقایسه و رتبه‌بندی تارنما[۱]های عمومی توجه زیادی را در روزهای اخیر به خود جلب کرده است. تبلیغ‌کنندگان، آژانس‌های تبلیغاتی، پژوهش‌گران دانشگاهی و مصرف‌کنندگان، علاقه وافری به رتبهٔ تارنماها نشان می‌دهند. دلیل رتبه‌بندی تارنماها از دو دیدگاه قابل بررسی است. از دیدگاه ارائه‌دهندگان خدمات، این نیاز وجود دارد که بتوانند مخاطبان خود را از تبلیغات بر روی سایت‌هایی که به‌وسیلهٔ تعداد زیادی از مشتریان بالقوه قابل مشاهده است، جذب کنند. از دیدگاه مصرف‌کنندگان، نیز این نیاز وجود دارد که بتوانند معتبرترین منابع را با توجه به جستجوی خود به‌دست بیاورند. به‌طورکلی اعتقاد بر این است که آنهایی که رتبه بالا در "فهرست جستجو" را از یک موتور جستجو و یا یک سایت پورتال به خود اختصاص می‌دهند به احتمال زیاد توسط جستجوگران اطلاعات، قابل مشاهده هستند و ترافیک بالای تارنما، پیش‌شرط خوبی برای انجام معاملات پرفایده است. در تجارت الکترونیک، روش‌های مختلفی برای رتبه‌بندی تارنماها به کار گرفته شده است. این روش‌ها را می‌توان با توجه به معیار مربوطه، به سه دسته گروه‌بندی کرد: معیارهای مبتنی بر فعالیت[۲]، معیارهای مبتنی بر مرجع[۳]، و معیارهای مبتنی بر نظر[۴].

در معیار مبتنی بر مرجع، سایت‌ها با توجه به تعداد دفعات بازدید سایت‌های دیگر از آنها در یک حیطهٔ موضوعی خاص، رتبه‌بندی می‌شوند. به‌احتمال سایت‌هایی که بیشترین بازدید را به‌وسیلهٔ سایت‌های دیگر به‌خصوص آنهایی که به‌عنوان یک مرجع معتبر در یک حیطهٔ موضوعی در نظرگرفته می‌شود، داشته باشند، اهمیت پیوندهای آن بیشتر می‌شود.

معیار مبتنی بر نظر در امر رتبه‌بندی تارنماها از نظرات هیئت داوران استفاده می‌کند. رتبه‌بندی‌های به‌دست‌آمده، بازتابی از داوری‌های ذهنی افراد هیئت مدیره هستند. تاحدودی تمام روش‌های رتبه‌بندی، از جمله روش‌های مبتنی بر ترافیک، به هیئت داوران اعتماد می‌کنند. اما معیار مبتنی بر نظر به‌طور صریح و با توجه کمی به داده‌های عینی، به نظرات ذهنی داوران بستگی دارد.

از این سه گروه، معیار مبتنی بر فعالیت که اغلب با عنوان رتبه‌بندی مبتنی بر ترافیک شناخته می‌شود از بقیه معیارها، مشهورتر بوده و به‌طورمعمول به‌عنوان عینی‌ترین روش محسوب می‌شود. در این معیار، تارنماها با توجه به میزان فعالیت‌هایی که در سایت اتفاق می‌افتد، رتبه‌بندی می‌شوند. سایت‌هایی که بیشترین ترافیک را جذب کنند و یا دارای بالاترین استفاده باشند در رتبه بالایی قرار می‌گیرند [2].

رتبه‌بندی ترافیکی[۵]، علی‌رغم اهمیت، تحت تأثیر ترافیک تقلبی می‌تواند آمار نادرستی از بازدیدکننده‌های یک سایت بدهد. ترافیک تقلبی شامل موارد زیر است [3]:

- استفاده از برنامه‌های پرداخت به‌ازای کلیک (PPC)[۶]
- ترافیک از سیستم‌های تبادل لینک / بازدیدکننده
- تبلیغات دامنه در انجمن‌ها، وبلاگ‌ها، چت روم، یا تارنماهای شبکه مانند فیس‌بوک
- تشویق خانواده و دوستان برای بازدید از صفحه به‌صورت توده‌ای
- پرداختن به ترویج دامنه با سرویس‌هایی مانند Google Adwords
- استفاده از پاپ آپ[۷]

[۱] Web site
[2] Activity_Based Criteria
[3] Reference_Based Criteria
[4] Opinion_Based Criteria
[5] Traffic Ranking
[6] Pay per click
[7] Pop up

- ایجاد ترافیک از طریق اسکریپت[1]ها، کلیک‌های ربات و غیره
- کلیک‌ها و ترافیک خود تولید[2] (کلیلک بر روی تبلیغات خود)

یکی از مهم‌ترین چالش‌های رتبه‌بندی، مقابله با ترافیک جعلی ایجادشده به‌وسیلهٔ برنامه‌های خودکار (ربات‌ها) است. ربات‌ها اجزای نرم‌افزاری مخرب مورد استفاده برای تولید هرزنامه‌ها، راه‌اندازی حملات مختل‌کنندهٔ سامانه[3]، فیشینگ[4]، سرقت هویت و خروج اطلاعات و دیگر فعالیت‌های غیر قانونی هستند [4]. تمایز بین ربات‌های وب و انسان‌ها به شرکت‌های بازاریابی کمک خواهد کرد تا اطلاعات آماری دقیق‌تری را در مورد تأثیر تبلیغات برخط و تعامل مشتریان واقعی با سایت‌های تجارت الکترونیک، به‌دست بیاورند. همچنین این تمایز به مدیران وب در تخمین اثرات جانبی واقعی فعالیت خزش گر در عملکرد وب سرور، کمک خواهد کرد. درنهایت، می‌تواند مبنایی برای توسعه سامانه‌های کنترل ورود هوشمند باشد که تارنماها را از خزش‌گرهای[5] مهاجم یا ناخواسته محافظت می‌کند [5].

تاکنون روش‌های مختلفی برای شناسایی ربات وب ارائه شده است. به‌طورکلی این روش‌ها را می‌توان به دو دستهٔ روش‌های برون‌خط و بلادرنگ تقسیم‌بندی کرد. روش بلادرنگ زمانی مهم است که موضوع نگران‌کننده، امنیت یک سایت باشد؛ درحالی‌که، روش برون‌خط برای فیلترکردن ترافیک ربات‌ها مورد استفاده قرار می‌گیرد [6]. در روش برون‌خط، از تجزیه و تحلیل لاگ دسترسی وب‌سرور[6] برای شناسایی و کشف ربات وب، استفاده می‌شود.

در این مقاله، روشی برای شناسایی ربات‌ها به‌منظور بهبود رتبه‌بندی ترافیکی تارنماها و برمبنای تجزیه و تحلیل لاگ دسترسی وب سرور، ارائه شده است. در این روش، نشست‌های موجود در لاگ شناسایی شده و سپس ویژگی‌هایی برای دسته‌بندی نشست‌ها به دو دسته انسان و ربات، استخراج می‌شود؛ سپس این نشست‌ها اغلب با سه شرط برچسب‌گذاری شده و درنهایت با الگوریتم‌های داده‌کاوی[7]، ربات‌های وب شناسایی می‌شود. نتایج به‌دست‌آمده نشان می‌دهد، روش پیشنهادی این پژوهش که به همراه استخراج ویژگی‌های جدید و به‌کارگیری شرط جدید در برچسب‌گذاری نشست‌ها است، باعث افزایش دقت در شناسایی ربات‌ها می‌شود.

ادامه این مقاله به این صورت سازماندهی می‌شود: در بخش ۲، کارهای انجام‌شده در گذشته و در زمینهٔ شناسایی ربات معرفی می‌شود. در بخش ۳، راه‌کار پیشنهادی این مقاله برای شناسایی ربات ارائه می‌شود. در بخش۴، به طراحی شرایط آزمون برای شناسایی ربات می‌پردازیم. در بخش ۵، آزمایش‌های انجام‌شده و نتایج آنها مورد بررسی قرار می‌گیرد؛ درنهایت نتیجه‌گیری و کارهای آینده در بخش ۶ معرفی می‌شود.

۲- کارهای مرتبط

پژوهش‌های مختلفی در زمینهٔ شناسایی ربات‌های ایجادکنندهٔ ترافیک برپایهٔ تجزیه و تحلیل لاگ دسترسی وب سرور انجام شده است. در نخستین مطالعات در سال ۲۰۰۲ *تان* و *کومار* الگوهای پیمایشی ربات‌های مختلف و انسان‌ها را بررسی کردند و ۲۵ ویژگی[8] از نشست‌ها[9] استخراج کردند که برای تمایز بین انسان از ربات مورد استفاده قرار می‌گیرد. آنها از الگوریتم درخت تصمیم c4.5 برای دسته‌بندی نشست‌ها استفاده کرده‌اند. نتایج آزمایش آنها نشان داد که ربات‌های وب می‌توانند با دقت بیش از ۹۰٪ بعد از چهار درخواست، شناسایی شوند [7]. *بومهارت* و همکارانش در سال ۲۰۰۵ ابزار پیش‌پردازش لاگ وب به نام RDT را توسعه دادند و از مدل شبکه‌های عصبی و رگرسیون منطقی برای تشخیص ربات وب استفاده و نتایج حاصل از کار خود را با نتایج به‌دست‌آمدهٔ تان و کومار مقایسه کردند. آنها در پژوهش خود علاوه‌بر استفاده از ۲۵ ویژگی کشف‌شده توسط تان و کومار، به معرفی ویژگی‌های جدید نیز پرداختند و درنهایت از ۳۴ ویژگی برای تمایز بین انسان و ربات استفاده کرده‌اند [8]. در سال ۲۰۰۹، *استسپلو* و همکارش از یک رویکرد بیزین[10] برای شناسایی خزش‌گرهای وب استفاده و نتایج به‌دست‌آمده از کار خود را با تکنیک درخت تصمیم مقایسه کرده‌اند. آنها در این مطالعه از ۶ ویژگی استخراج‌شده از لاگ دسترسی وب‌سرور استفاده کردند [5]. در مطالعه‌ای که در سال ۲۰۱۲ توسط *استیوانوویچ* و همکارانش انجام شد، از هفت الگوریتم یادگیری با نظارت به منظور دسته‌بندی نشست‌های کاربران به دو دستهٔ خزش‌گرهای خودکار وب و بازدیدکنندگان انسانی، استفاده شده بود. آنها در روش خود، به معرفی دو ویژگی

[1] script
[2] Self_generated
[3] Distributed Denial of Service(DDOS)
[4] Phishing
[5] Crawler
[6] Web Server Access Log
[7] Data Mining
[8] Feature
[9] Session
[10] Bayesian

جدید نیز پرداخته و درنهایت از هشت ویژگی در کار خود بهره بردند. نتایج تجربی قدرت دو ویژگی جدید در بهبود الگوریتم‌های رده‌بندی برای شناسایی خزش‌گرهای مخرب و خوش رفتار را نشان داده است [9]. همچنین در سال ۲۰۱۳ استیوانوویچ و همکارانش از الگوریتم‌های یادگیری بدون نظارت شبکهٔ عصبی به نام نقشهٔ خودسازمان‌ده[1] (SOM) و نظریهٔ تشدید انطباقی دو[2] (ART2) برای شناسایی ربات وب استفاده کرده‌اند. آن‌ها در این پژوهش به معرفی دو ویژگی جدید دیگر نیز پرداختند و در مجموع از ده ویژگی استفاده کردند [10]. *رجب‌نیا* و همکارانش در سال ۲۰۱۳ از منطق استنتاج فازی و مبتنی بر درخت تصمیم در امر شناسایی ربات استفاده کرده‌اند. آنها از چهارده ویژگی برای تمایز بین انسان و ربات بهره برده‌اند و نتایج الگوریتم پیشنهادی خود را با روش‌های یادگیری وزن‌دارشدهٔ محلی، طبقه‌بندی‌کننده Adaboost و Bagging، C4.5، شبکه بیزین و شبکه باور بیزی مقایسه کرده‌اند. نتایج محاسبهٔ آنها نشان داد که روش پیشنهادیشان دقت بالاتری نسبت به روش‌های مورد مقایسه دارد [1]. در سال ۲۰۱۴، *وفایی‌جهان* و همکارانش از روش خوشه‌بندی مبتنی بر چگالی برای شناسایی ربات وب استفاده کرده‌اند. آنها در این مطالعه به معرفی دو ویژگی جدید نسبت به پژوهش‌های انجام‌شده در این زمینه پرداختند. در این پژوهش از الگوریتم DBSCAN برای خوشه‌بندی بازدیدکنندگان وب استفاده شده است. به‌دلیل مشکل بعدپذیری الگوریتم پیشنهادی، تنها چهار ویژگی از نشست‌ها استخراج شده است. نتایج مطالعات آنها نشان داد که از نقطه نظر دقت و کیفیت خوشه‌بندی، روش پیشنهادی بهتر از الگوریتم‌های قبل عمل می‌کند[11]. *سیسودیا* و همکارانش در سال ۲۰۱۵ اثربخشی یادگیرنده‌های مبتنی بر گروه را در شناسایی نشست‌های ربات وب از لاگ دسترسی وب‌سرور، مورد بررسی قرار داده‌اند. در این پژوهش، یک مطالعه مقایسه‌ای از نظر دسته‌بندی بین توابع گروهی (Bagging, Boosting, Voting) و دسته‌بندی‌کننده‌های ساده انجام شده و اثربخشی این دسته‌بندی‌کننده‌ها (گروهی و ساده) را بر روی پنج مجموعه‌دادهٔ متفاوت از نظر طول نشست، ارزیابی شده است. آنها از ۲۳ ویژگی در کار خود بهره بردند [12]. تا اینجا در زمینه تشخیص ربات وب، مطالعات معتبری به معرفی ویژگی‌های جدید برای شناسایی نشست‌های وب اختصاص داده شده است. با این حال، ماهیت متفاوت تارنماها و رفتار تغییرپذیر بازدیدکنندگان وب می‌تواند تأثیرات قابل توجهی را بر روی اثربخشی ویژگی‌های انتخاب‌شده برای معرفی نشست‌های یک تارنما، بگذارد. برای مقابله با این ابهام ذاتی حال حاضر داده‌های وب، در سال ۲۰۱۷ *حمیدزاده* و همکارانش، از یک الگوریتم جدید به نام FRS_WRD که مبتنی بر تئوری مجموعه سخت فازی است برای رسیدگی به داده‌های مبهم و نویزی و انتخاب مرتبط‌ترین ویژگی برای نشست‌های وب، استفاده کرده‌اند. در حقیقت خصوصیت جدید FRS_WRD این است که به‌صورت پویا می‌تواند ویژگی‌های استفاده شده برای معرفی بازدیدکنندگان وب را نه‌تنها برای افزایش عملکرد تشخیص، بلکه همچنین برای حذف مشکل چندبعدی، انتخاب کند. در این مطالعه از الگوریتم SOM شبکهٔ عصبی برای خوشه‌بندی بازدیدکنندگان وب استفاده شده است. نتایج کار آنها نشان می‌دهد که برخلاف ویژگی‌های قابل قبولی که در پژوهش‌های گذشته معرفی شده است، انتخاب پویای ویژگی‌های هر مجموعه‌داده در عملکرد نهایی بسیار مؤثر است. همچنین ارزیابی‌های انجام‌شده نشان داده است که به‌طورکلی در مقایسه با الگوریتم‌های قبلی، FRS_WRD نتایج بهتری را به‌دست آورده است [13].

در پژوهش‌های یادشده در بالا، ویژگی‌های استخراج‌شده از هر نشست برای تمایز بین انسان و ربات به کار برده شده است؛ درحالی‌که، روش پیشنهادی این مقاله، با دید کلی‌تری به استخراج ویژگی از هر کاربر می‌پردازد که درنتیجه باعث افزایش دقت و سرعت در شناسایی ربات می‌شود.

در مقایسه با مطالعات انجام‌گرفته در گذشته، نوآوری‌های این مقاله شامل موارد زیر است:

۱. این مقاله با دید بهبود رتبه‌بندی ترافیکی، کشف تقلب ایجادشده به‌وسیلهٔ ربات‌ها را مورد بررسی قرار می‌دهد.
۲. در این مقاله، از الگوریتم‌های دسته‌بندی برای شناسایی ربات وب استفاده می‌کنیم. نتایج حاصل از پژوهش نشان می‌دهد که روش پیشنهادی ما با کشف ویژگی‌های جدید از رفتار هر کاربر و معرفی شرط جدید در برچسب گذاری نشست‌ها، با دقت بالاتری ربات‌های ایجادکنندهٔ ترافیک را شناسایی می‌کند.

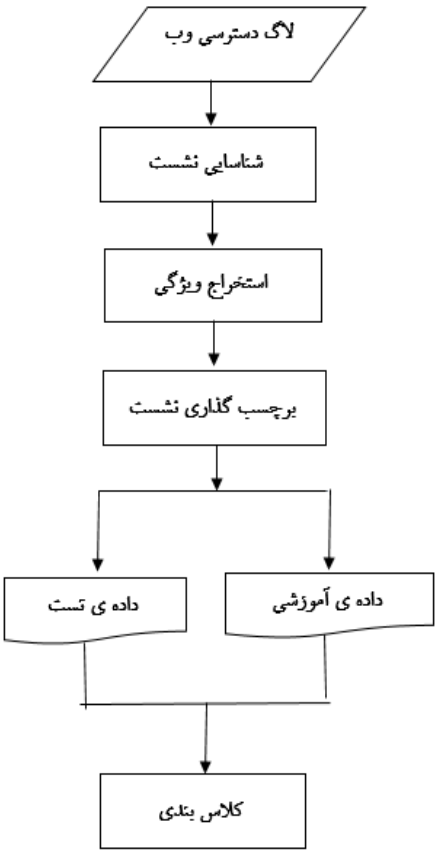
۳- راه‌کار پیشنهادی برای شناسایی ربات

[1] Self Organizing Map
[2] Adaptive Resonance Theory 2

در این مقاله تلاش می‌شود تا ترافیک ایجادشده به‌وسیلهٔ ربات‌ها بر مبنای درخواست‌های ذخیره‌شده در لاگ دسترسی وب‌سرور، شناسایی شود. در این پژوهش، تجزیه و تحلیل‌کنندهٔ لاگ مبتنی بر C# بوده و فعالیت‌های زیر را انجام می‌دهد:

۱. پویش‌کردن ورودی‌های لاگ برای شناسایی نشست‌های[1] بازدیدکنندگان
۲. استخراج ویژگی[2] از نشست‌های کاربر
۳. برچسب‌گذاری نشست‌ها[3]
۴. دسته‌بندی نشست‌ها به دو دستهٔ ربات یا انسان

مراحل اصلی در روش پیشنهادی، برطبق روندنمای شکل (۱) به شرح زیر می‌باشد:



(شکل-۱): روندنمای شناسایی ربات
(Figure-1): Robot detection flowchart

۳-۱- لاگ دسترسی وب‌سرور

لاگ دسترسی وب‌سرور اطلاعات جزئی از هر درخواست تولیدشده از مرورگر وب کاربر به سمت وب سرور را به‌ترتیب زمانی ذخیره می‌کند. لاگ‌ها به‌طور معمول دارای اطلاعاتی از قبیل برچسب زمانی، نشانی IP کاربر درخواست‌کننده، نوع تابع درخواست، عامل کاربری، نوع فایل درخواستی و غیره هستند. یک نمونه لاگ دسترسی ساده به‌صورت زیر است. توضیحات هر فیلد در جدول (۱) آمده است.

(جدول-۱): معرفی فیلدهای موجود در لاگ
(Table-1): Introduce the fields in the log

ردیف	فیلد	شرح فیلد
1	127.0.0.1	نام میزبان یا آدرس IP کاربری که درخواستی را به سرور می‌فرستد.
2	-	اطلاعات درخواست‌شده را نشان می‌دهد.
3	Frank	شناسهٔ کاربر احراز هویت شده را نشان می‌دهد.
4	[10/Oct/2000:13:55:36 - 0700]	زمان را به فرمت رایج لاگ نشان می‌دهد.
5	"GET /apache_pb.gif HTTP/1.0"	درخواست فرستاده شده از سمت کاربر و تابع درخواست را نشان می‌دهد.
6	200	کد وضعیت فرستاده شده توسط سرور را نشان می‌دهد.
7	2326	بایت‌های فرستاده شده از سمت سرور به مشتری را نشان می‌دهد.
8	"http://www.example. Com/start html"	سایتی را که گزارش‌های مشتری از آن ارجاع داده شده است را نشان می‌دهد.
9	"Mozilla/4.08 [en] (Win98; I ;Nav)"	اطلاعاتی را که مرورگر مشتری در مورد خود به سرور گزارش می‌کند را نشان می‌دهد.

۳-۲- شناسایی نشست

در این مرحله، کلیهٔ درخواست‌های HTTP براساس آدرس IP و عامل کاربری یکسان گروه‌بندی می‌شوند. در این قسمت از یک رویکرد برای شکستن یک گروه به زیرگروه‌های دیگر استفاده می‌شود به این صورت که اگر زمان بین دو درخواست از یک زیرگروه IP بیشتر از یک حد آستانه باشد، این طور فرض می‌شود که کاربر یک نشست جدید را آغاز کرده است. اغلب حد آستانه در بسیاری از مطالعات ۳۰ دقیقه در نظر گرفته می‌شود [4].

[1] Session Identification
[2] Feature Extraction
[3] Session labeling

۳-۳- استخراج ویژگی

برای تمام نشست‌های به‌دست‌آمده، ویژگی‌هایی از درخواست‌های موجود در آن نشست استخراج می‌شود که برای تمایز بین انسان و ربات مورد استفاده قرار می‌گیرد. در این مقاله از ۲۸ ویژگی برای شناسایی ربات استفاده شده است. ویژگی‌های ۱۹ تا ۲۸، ویژگی‌های جدیدی هستند که در این مقاله معرفی شده‌اند. ویژگی‌های استخراج شده به‌صورت زیر می‌باشند:

۱. زمان شروع نشست: زمان شروع اولین درخواست در نشست

۲. طول نشست: به مدت زمان بین نخستین و آخرین درخواست گفته می‌شود. هرچه مدت زمان بیشتر باشد؛ بازدیدکننده به ربات نزدیک‌تر است [5].

۳. تعداد درخواست‌ها: تعداد کل درخواست‌های یک نشست را محاسبه می‌کند [9] .

۴. درخواست فایل HTML: تعداد درخواست‌های فایل از نوع HTML را نشان می‌دهد. پسوند فایل HTML :.html.php, .htm, .js, .cgi etc [12].

۵. درخواست فایل تصویر (Image): تعداد درخواست‌های از نوع فایل تصویری با پسوندهایی از قبیل .jpg, .png,.jpeg .tiff, .ico, .gif etc را بررسی می‌کند [12].

۶. نسبت HTML به تصویر: تعداد درخواست صفحات HTML تقسیم بر تعداد درخواست صفحات تصویر را در یک نشست محاسبه می‌کند. ربات‌های وب نسبت به کاربران انسانی فایل HTML بیشتری را درخواست می‌دهند، درصورتی که، کاربران انسانی، درخواست بیشتری برای تصویر دارند؛ بنابراین هر چه این نسبت بیشتر باشد، بازدیدکننده به ربات نزدیک‌تر است [10].

۷. درخواست از نوع Head: تعداد درخواست‌های فرستاده‌شده با تابع head را محاسبه می‌کند.

۸. درصد درخواست فایل با تابع Head: درصد درخواست از نوع head را نسبت به کل درخواست‌ها محاسبه می‌کند. ربات‌های وب برای کاهش حجم اطلاعات مربوط به درخواست، از ارسال یک عنوان به سرور استفاده می‌کنند. این درحالی‌است که کاربران انسانی با استفاده از مرورگرها درخواست خود را به سرور ارسال می‌کنند، که در این روش، تابع درخواست GET است [10].

۹. درخواست با ارجاع انتساب داده‌نشده: تعداد درخواست‌ها با پیوند ارجاع خالی را محاسبه می‌کند.

۱۰. درصد درخواست با ارجاع انتساب داده نشده: درصد درخواست‌ها با ارجاعات خالی را نسبت به کل درخواست‌ها در یک نشست نشان می‌دهد. این ویژگی برای بیش‌تر ربات‌ها مقدار بالایی دارد چون مرورگرهای وب اغلب اطلاعاتی را به‌عنوان ارجاعات و به‌صورت پیش‌فرض مقداردهی می‌کنند، اما در ربات‌ها این فیلد به‌طورعمومی مقدار "-" می‌گیرد [10].

۱۱. درخواست با کد خطای 4xx: تعداد درخواست‌ها در یک نشست که شامل کد خطای 4xx هستند را محاسبه می‌کند.

۱۲. درصد درخواست با کد خطای 4xx: درصد درخواست‌ها با این کد خطا را تقسیم بر کل درخواست‌های موجود در آن نشست، نشان می‌دهد. ربات‌های وب نسبت به کاربران انسانی با درصد بالاتری پیوندهای خراب را انتخاب می‌کنند [5].

۱۳. انحراف استاندارد از عمق صفحات درخواست‌شده: یک خصیصه عددی است که انحراف استاندارد از عمق صفحه در تمام درخواست‌های موجود در یک نشست را نشان می‌دهد. برای مثال، درخواست " google/translate/persianpage.html/ " با عمق سه و درخواست " google/translate.html/ " با عمق دو در نظر گرفته می‌شود. در یک مرور معمولی وب، انسان‌ها برای پیداکردن اطلاعات مورد علاقهٔ خود، یک مجموعه از پیوندهای مرتبط و به هم پیوسته و خاص‌تر را دنبال می‌کنند. در مقابل، ربات‌ها چنین الگوی پیمایشی پیچیده‌ای ندارند و نمی‌توانند به‌وسیلهٔ ساختار پیوند تارنما محدود شوند. به‌طور مثال، از آنجایی که خزشگرها مانند گوگل، کل دامنهٔ یک وب را پیمایش می‌کنند، می‌توانند فایل‌ها را در عمق‌های مختلف در سلسله‌مراتب فایل‌ها پیدا کنند و درنتیجه انحراف استاندارد مقدار بیشتری به خود می‌گیرد. از سوی دیگر انسان‌ها تمایل دارند بر روی یک نوع خاص از اطلاعات که به‌طورمعمول در فایل‌های کمتری در یک دایرکتوری تنها، ذخیره می‌شوند، تمرکز کنند. در نتیجه، با توجه به دلایل یادشده، انحراف استاندارد از عمق صفحات برای ربات‌ها مقدار بیشتری نسبت به کاربران انسانی خواهد داشت [9].

۱۴. تعداد بایت‌های درخواست‌شده از سرور: مقدار داده به‌صورت بایت که در یک نشست واحد از سرور درخواست شده است، را نشان می‌دهد. به‌طورعمومی ربات‌های وب مقدار داده‌های بیشتری

را نسبت به بازدیدکنندگان انسانی در یک نشست واحد از سرور درخواست می‌کنند [1].

۱۵. درخواست فایل CSS: تعداد درخواست‌ها از نوع فایل .css را نشان می‌دهد.

۱۶. درصد درخواست فایل CSS: مرورگرهای وب به‌صورت خودکار یک درخواست برای فایل css ارسال می‌کنند درحالی‌که ربات‌های وب نیازی به مشاهده فایل css ندارند. پس اگر در یک نشست، تمام درخواست‌ها از یک نوع منبع باشد و فایل css مشاهده نشده باشد، آن‌گاه آن نشست مربوط به ربات خواهد بود [1].

۱۷. درخواست فایل PDF/PS: تعداد درخواست‌ها از نوع فایل ps. یا pdf. را نشان می‌دهد.

۱۸. درصد درخواست فایل PDF/PS: درصد درخواست فایل‌های PDF یا PS را در یک نشست محاسبه می‌کند. ربات‌های وب درخواست‌های بیشتری برای فایل‌های PDF و Postscript جهت جمع‌آوری اطلاعات دارند [10].

۱۹. تعداد نشست‌های کاربر: تمام نشست‌هایی را که یک کاربر در فایل لاگ دسترسی وب‌سرور در آن حضور دارد محاسبه می‌کند. ربات‌های وب برای ایجاد ترافیک جعلی ممکن است، نشست‌های بیشتری در زمان‌های کوتاه‌تری ایجاد کنند؛ درنتیجه هرچه تعداد نشست‌ها بیشتر باشد، کاربر به ربات نزدیک‌تر است.

۲۰. طول نشست‌های یک کاربر منحصر به فرد: طول هر نشست را برای هر کاربر با توجه به تعداد نشست‌های آن کاربر، محاسبه می‌کند. نشست‌های یک خزش‌گر زمان طولانی‌تری را نسبت به نشست‌های انسان سپری می‌کند. رفتار پیمایش وب انسان‌ها متمرکزتر و هدفمندتر از ربات‌های وب است. درنتیجه حد مشخصی از زمانی که یک انسان در سایتی پیمایش می‌کند، وجود دارد [5]. پس هر چه این عدد بزرگ‌تر باشد، بازدیدکننده به ربات نزدیک‌تر است.

۲۱. میانگین طول نشست‌های یک کاربر منحصربه‌فرد.

۲۲. واریانس طول نشست‌های یک کاربر منحصربه‌فرد.

۲۳. تعداد درخواست‌های یک کاربر منحصربه‌فرد: تمام درخواست‌های یک کاربر را در تمامی نشست‌هایی که آن کاربر در آن حضور دارد، اندازه‌گیری می‌کند.

۲۴. نرخ خطای 4xx برای یک کاربر منحصربه‌فرد: تمامی درخواست‌های کاربر را در تمامی نشست‌هایی که شرکت کرده و با پاسخ 4xx مواجه شده است، اندازه‌گیری می‌کند. بالابودن مقدار این ویژگی، نشان‌دهندهٔ ربات است.

۲۵. میانگین نرخ خطای 4xx برای یک کاربر منحصربه‌فرد.

۲۶. شناسه نشست: عددی که به‌منظور شناسایی به هر نشست اختصاص داده می‌شود.

۲۷. شناسه کاربر: عددی که به‌منظور شناسایی به هر کاربر اختصاص داده می‌شود.

۲۸. روزها: بازهٔ زمانی یک‌روزه را برای محاسبهٔ نشست‌های یک کاربر نشان می‌دهد.

۴-۳- برچسب‌گذاری نشست

در این مرحله، ابتدا تمام نشست‌ها را به‌صورت پیش‌فرض انسان در نظر می‌گیریم مگر اینکه یکی از سه شرط زیر در آن صدق کند:

- اگر نشانی IP نشست در فهرست نشانی IPهای ربات‌های شناخته‌شده وجود داشته باشد، آن نشست به‌عنوان ربات برچسب‌گذاری می‌شود [4].
- اگر عامل کاربری نشست در فهرست عامل کاربری ربات‌های شناخته‌شده موجود باشد، آن نشست ربات در نظر گرفته می‌شود [4].
- درصورتی‌که درخواستی از یک نشست شامل فایل Robots.txt باشد، آن نشست نیز به‌عنوان ربات برچسب‌گذاری می‌شود [1].

با توجه به این‌که برخی از ویژگی‌های استخراج‌شده بر مبنای هر کاربر است می‌توان این شرط جدید در برچسب گذاری نشست را در نظر گرفت که: اگر یک نشست از یک کاربر منحصربه‌فرد، برچسب ربات گرفت، بقیهٔ نشست‌های آن کاربر هم ربات شناخته شود (این شرط در آزمایش آخر مورد استفاده قرار می‌گیرد).

۵-۳- دسته‌بندی

دسته‌بندی یکی از روش‌های یادگیری با نظارت بوده و به این صورت است که ابتدا در مرحلهٔ آموزشی، مدل مورد نظر ساخته می‌شود و سپس در مرحلهٔ ارزیابی، کارایی مدل آزمایش می‌شود. در این پژوهش برای دسته‌بندی کاربران به دو دستهٔ انسان و ربات، از پنج الگوریتم دسته‌بندی ID3, Random Forest, Naïve Bayes, Hidden Naïve Bayes, Bayesian net استفاده می‌کنیم.

۴- طراحی شرایط آزمون

لاگ استفاده شده در این پژوهش از تاریخ ۲۰۱۲/۰۳/۱۶ تا تاریخ ۲۰۱۲ /۰۳/۱۷ و از سایت www.secrepo.com

جمع‌آوری شده است [17]. این لاگ شامل ۲،۰۴۸،۴۴۳ رکورد است. اطلاعات موجود در این لاگ در جدول (۲) آورده شده است.

برای دسته‌بندی کاربران به دو دستهٔ انسان و ربات، از الگوریتم‌های دسته‌بندی موجود در نرم‌افزار وکا[1] بهره می‌بریم. معیار ارزیابی الگوریتم‌های رده‌بندی در این مقاله، دقت حاصل از شناسایی ربات است. همان‌طور که در جدول (۲) مشاهده می‌شود، یک عدم بالانس بین کلاس‌های انسان و ربات وجود دارد. درنتیجه، دقت کلی حاصل از دسته‌بندی، مقدار غیر واقعی به خود می‌گیرد. از آنجایی که هدف این پژوهش شناسایی ربات‌های ایجادکنندهٔ ترافیک است، به همین دلیل، در آزمایش ۱ و۲، تنها دقت شناسایی ربات را مدنظر قرار داده و این دقت را تنها در دادهٔ آزمایش بررسی کرده و روش پیشنهادی را براساس آن ارزیابی می‌کنیم. این دقت از تقسیم تعداد نمونه‌هایی که صحیح دسته‌بندی شده‌اند به کل نمونه‌ها، به‌دست می‌آید. برای به‌دست‌آوردن این دو مقدار، از ماتریس درهم‌ریختگی[2] حاصل از دسته‌بندی استفاده می‌کنیم. چهار پارامتر زیر را با توجه به این ماتریس تعریف می‌کنیم:

- TP: نشست‌هایی که ربات برچسب‌گذاری‌شده و الگوریتم دسته‌بندی هم صحیح پیش‌بینی کرده است.
- FN: نشست‌هایی که ربات برچسب‌گذاری شده‌اند، ولی الگوریتم دسته‌بندی، اشتباه پیش‌بینی کرده است.
- FP: نشست‌هایی که انسان برچسب‌گذاری شده‌اند ولی الگوریتم دسته‌بندی، اشتباه پیش‌بینی کرده است.
- TN: نشست‌هایی که انسان برچسب‌گذاری شده‌اند و الگوریتم دسته‌بندی هم صحیح پیش‌بینی کرده است.

با توجه به پارامترهای یادشده، دقت شناسایی ربات به‌صورت زیر محاسبه می‌شود:

$$Accuracy = {TP}/{TP + FN} \qquad (۱)$$

به منظور دسته‌بندی، ابتدا از ۷۰٪ داده‌ها به‌عنوان دادهٔ آموزش استفاده کرده و مدل مورد نظر را می‌سازیم. سپس با ۳۰٪ داده‌های باقی‌مانده، مدل آزمایش می‌شود. اندازه‌گیری دقت در قسمت آزمایش و ارزیابی مدل بررسی می‌شود.

(جدول-۲): معرفی مجموعهٔ داده
(Table-2): Introducing the data set

تعداد کل نمونه‌ها	**289257**
تعداد نشست‌های انسان	288795
تعداد نشست‌های ربات	462
تعداد نمونه‌ها در مجموعه‌ی آموزش	202479
تعداد نمونه‌ها در مجموعهٔ آزمایش	86778

۵- آزمایش و ارزیابی مدل پیشنهادی

در این مقاله، به منظور بررسی و ارزیابی مدل پیشنهادی، سه آزمایش در نظر گرفته شده است. در آزمایش نخست، تنها از ویژگی‌های استفاده شده در مطالعات قبل که در بخش سوم معرفی شد، استفاده می‌گردد. در آزمایش دوم، ویژگی‌های قبلی به‌علاوهٔ ویژگی‌های جدید پیشنهادی در این پژوهش، مورد استفاده قرار می‌گیرد و دقت حاصل از این دو آزمایش با هم مقایسه می‌شوند؛ درنهایت در آزمایش سوم، شرط جدید در برچسب‌گذاری نشست‌ها را اعمال کرده و نتایج مورد بررسی قرار می‌گیرند.

۵-۱- آزمایش اول: ارزیابی دقت دسته‌بندی با استفاده از ویژگی‌های قبلی

همان‌طور که گفته شد، منظور از دقت مورد بررسی در این مقاله، دقت حاصل از شناسایی ربات در دادهٔ آزمایش است. دادهٔ آزمایش دارای ۸۶۷۷۸ نمونه است که از این تعداد ۱۴۰ نمونه برچسب ربات و ۸۶۶۳۶ نمونه برچسب انسان به خود گرفته‌اند. در این آزمایش از ۱۹ ویژگی معرفی شده در مقالات قبل، استفاده می‌شود. دقت حاصل از رده‌بندی در جدول زیر آمده است:

(جدول-۳): نتایج حاصل از ارزیابی دقت دسته‌بندی با استفاده از ویژگی‌های قبلی (آزمایش نخست)
(Table-3): Results of evaluation of classification accuracy using previous features (first experiment)

نوع الگوریتم	تعداد نمونه‌های صحیح دسته‌بندی شده/ کل نمونه‌ها	٪دقت
Naïve Bayes	$^{14}/_{140}$	10
Bayes Net	$^{14}/_{140}$	10
HNB	$^{11}/_{140}$	8.7
Id3	$^{14}/_{140}$	10
Random Forest	$^{27}/_{140}$	10

۵-۲- آزمایش دوم: ارزیابی دقت دسته‌بندی با استفاده از ویژگی‌های جدید پیشنهادی

با توجه به یکسان‌بودن مجموعه‌داده، تعداد نمونه‌های دادهٔ آزمایش و تعداد ربات‌ها و انسان‌ها در این مجموعه‌داده مانند آزمایش نخست است. در این آزمایش علاوه‌بر استفاده از نوزده ویژگی آزمایش قبل، از نه ویژگی جدید

[1] Weka
[2] Confusion Matrix

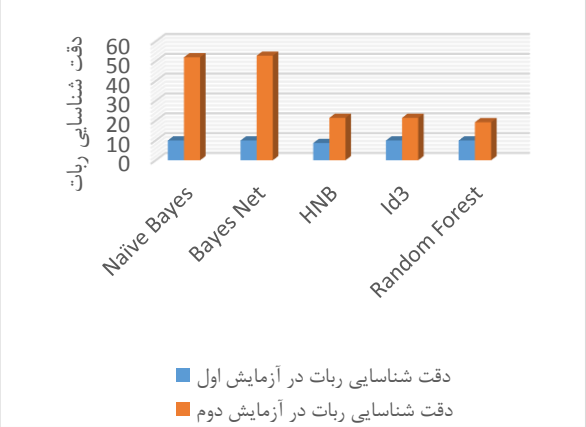
استخراج‌شده نیز بهره می‌بریم. به‌صورت کلی تعداد ویژگی‌های مورد استفاده در این آزمایش که برای تمایز بین انسان و ربات به کار برده می‌شود، ۲۸ عدد است. نتایج حاصل از آزمایش در جدول (۴) آورده شده است:

(جدول-۴): نتایج حاصل از ارزیابی دقت دسته‌بندی با استفاده از ویژگی‌های جدید پیشنهادی (آزمایش دوم)

(Table-4): Results of evaluation of classification accuracy using the proposed new features (second experiment)

نوع الگوریتم	تعداد نمونه‌های صحیح دسته‌بندی‌شده/ کل نمونه‌ها	٪دقت
Naïve Bayes	$^{73}/_{140}$	52
Bayes Net	$^{74}/_{140}$	52.8
HNB	$^{30}/_{140}$	21.4
Id3	$^{30}/_{140}$	21.4
Random Forest	$^{27}/_{140}$	19.2

مقایسهٔ بین نتایج حاصل از دو آزمایش انجام‌شده در نمودار آورده شده است. این نمودار نشان می‌دهد که ویژگی‌های جدید پیشنهادشده در تعامل با ویژگی‌های قبلی، دقت به‌دست‌آمده را در تمامی الگوریتم‌های مورد بررسی، افزایش می‌دهد و درنتیجه باعث بهبود در شناسایی ربات ایجادکنندهٔ ترافیک می‌شود.



(نمودار-۱): نمودار مقایسه‌ای دقت بین آزمایش‌های نخست و دوم

(Chart-1): Accuracy comparison chart between the first and second experiments

۳-۵- ارزیابی ارزش ویژگی‌های پیشنهادی

برای اندازه‌گیری ارزش ویژگی‌های جدید پیشنهادشده، ۲۸ ویژگی به‌دست‌آمده را که در آزمایش‌های دوم و سوم مورد استفاده قرار گرفته و از ترکیب ویژگی‌های قدیم و جدید به‌دست آمده‌اند رتبه‌بندی می‌کنیم (ویژگی ۲۹، برچسب رده است و در رتبه‌بندی لحاظ نمی‌شود). برای رتبه‌بندی ویژگی‌ها از نرم‌افزار وکا استفاده می‌شود. در قسمت انتخاب ویژگی، ارزش یک ویژگی را از طریق اندازه‌گیری معیار IG با توجه به دسته (کلاس) به‌صورت زیر مشخص می‌کنیم [16]:

$$InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute) \quad (۲)$$

نتایج حاصل از رتبه‌بندی ویژگی‌ها در جدول (۵) آمده است:

(جدول-۵): رتبه‌بندی ویژگی‌ها

(Table-5): Features ranking

رتبه	ویژگی
1	Error4XXMean
2	Session Count
3	Start Epoch
4	Error4XXRatioPer
5	Session.ID
6	Session Duration Total
7	User.ID
8	No Ref Link Ratio Per
9	Head Request Ratio Per
10	SessionSet.Error4XXCount
11	SessionSet.Request Count
12	Html To Image Ratio Per
13	No Ref Link Count
14	HTML Request Count
15	Request Count
16	CSS Request Count
17	Head Request Count
18	Duration
19	Error4XXCount
20	Image Request Count
21	Session Duration Variance
22	CSS Request Ratio Per
23	Session Duration Mean
24	Depth Standard Deviation
25	PDF/PS Request Ratio Per
26	PDF/PS Request Count
27	Bytes Sent
28	Day

در جدول (۵)، ویژگی‌های جدید پیشنهادی به‌صورت پررنگ مشخص شده‌اند. همان‌طوری‌که در این جدول

مشاهده می‌شود، اغلب ویژگی‌های جدید پیشنهادشده، ارزش بالایی را در رتبه‌بندی ویژگی‌ها به‌دست آورده‌اند و لذا ویژگی‌های مناسبی هستند.

۴-۵- آزمایش سوم: ارزیابی دقت دسته‌بندی با به‌کارگیری شرط جدید در برچسب‌گذاری نشست‌ها

همان‌طور که گفته شد، می‌توانیم با توجه به ویژگی‌های جدید پیشنهاد شده، شرط جدیدی را نیز در برچسب‌گذاری نشست‌ها در نظر بگیریم به این صورت که اگر یک نشست از یک کاربر منحصربه‌فرد، برچسب ربات گرفت، بقیهٔ نشست‌های آن کاربر هم ربات شناخته شود. تعداد نمونه‌های به‌دست‌آمده در نتیجهٔ اجرای برنامه ۲۸۹۲۵۷ نمونه است. از این تعداد، ۲۰۲۴۷۹ نمونه در مجموعهٔ آموزش و ۸۶۷۷۸ نمونه در مجموعهٔ آزمون قرار دارند. تعداد نشست‌هایی که به‌عنوان انسان برچسب‌گذاری شده‌اند ۱۷۹۵۵۶ نشست بوده و تعداد ۱۰۹۷۰۱ نشست نیز ربات است. همان‌طور که مشاهده می‌شود، بین نشست های انسان و ربات، بالانس قابل قبولی وجود دارد. در نتیجه، دقت به‌صورت کلی و هم برای نشست‌های انسان و هم برای نشست‌های ربات اندازه‌گیری می‌شود. نتایج حاصل از این آزمایش در جدول (۶) آمده است:

(جدول-۶) : نتایج حاصل از ارزیابی دقت دسته‌بندی با به‌کارگیری شرط جدید در برچسب‌گذاری نشست‌ها (آزمایش سوم)

(Table-6): Results of Evaluating the Accuracy of Classification Using a New Condition in Session Labeling (Experiment 3)

الگوریتم	Correctly Classified	Incorrectly classified	Precision	Recall	F1
Naïve bayes (Train)	198048	4431	0.979	0.978	0.978
Naïve bayes (Test)	84854	1914	0.978	0.978	0.978
Bayes net (Train)	198053	4426	0.979	0.978	0.978
Bayes net (Test)	84865	1913	0.979	0.978	0.978
HNB (Train)	201197	1281	0.978	0.994	0.994
HNB (Test)	86239	539	0.994	0.994	0.994
ID3 (Train)	202159	320	0.994	0.998	0.998
ID3 (Test)	86610	153	0.998	0.998	0.998
Random Forest (Train)	202159	320	0.998	0.998	0.998
Random Forest (Test)	86625	153	0.998	0.998	0.998

همان‌طوری‌که در جدول (۶) مشاهده می‌شود، مدل پیشنهادی به همراه شرط جدید برچسب‌گذاری، دقت قابل قبولی را در شناسایی ربات به‌دست آورده به‌طوری که در برخی الگوریتم‌ها، این دقت مقدار بالای ۹۹ درصد را کسب کرده است.

۶- نتیجه‌گیری و کارهای آینده

بسیاری از صاحبان تجارت الکترونیکی به‌دنبال آمار واقعی بازدیدکنندگان از سایت‌هایشان هستند تا بدین صورت بتواند کسب‌وکار خود را بهبود ببخشند. معیاری به نام رتبه‌بندی ترافیکی برای این منظور استفاده می‌شود. رتبه‌بندی ترافیکی، میزان ترافیک جذب‌شده به سایت‌های کسب و کار و یا فعالیت‌های انجام‌شده در آن را اندازه‌گیری می‌کند. فریب‌های متفاوتی این رتبه‌بندی را مورد هدف قرار می‌دهند که یکی از آنها، ایجاد ترافیک جعلی توسط بازدیدکنندگان غیر انسانی (ربات‌ها) است. برای داشتن یک رتبه‌بندی صحیح و قابل اعتماد، لازم است تا ترافیک مربوط به ربات‌ها شناسایی و حذف شود.

در این پژوهش از ویژگی‌های جدیدی برای شناسایی ربات‌های ایجادکنندهٔ ترافیک، استفاده شده است. آزمایش‌ها نشان می‌دهد که طبق رتبه‌بندی به‌عمل‌آمده درخصوص ویژگی‌ها، اغلب ویژگی‌های پیشنهادی دارای ارزش بالایی هستند. همچنین استفاده از ویژگی‌های جدید پیشنهادشده در تعامل با ویژگی‌های قبلی، دقت شناسایی ربات ایجادکنندهٔ ترافیک را در تمامی الگوریتم‌های مورد بررسی، افزایش می‌دهد.

همچنین در این پژوهش از شرط جدیدی در برچسب‌گذاری نشست‌ها استفاده شده است. آزمایش‌ها نشان می‌دهد که مدل پیشنهادی به همراه افزودن شرط جدید در برچسب‌گذاری، دقت قابل قبولی را در شناسایی ربات به‌دست آورده به‌طوری که در بیش‌تر الگوریتم‌ها، این دقت مقدار بالای ۹۹ درصد را کسب کرده است. در نتیجه این افزایش دقت در شناسایی ربات‌ها و حذف ترافیک ساختگی ایجادشده توسط آنها، رتبه‌بندی ترافیکی تارنماها را بهبود داده و موجب حفظ عدالت در رتبه‌بندی می‌شود.

ما در این پژوهش، از الگوریتم‌های رده‌بندی استفاده کردیم. به‌عنوان پیشنهاد برای کارهای آینده می‌توان از الگوریتم‌های خوشه‌بندی برای تقسیم نشست‌ها به دو دستهٔ انسان و ربات استفاده کرد. همچنین می‌توان سایر روش‌های شناسایی ربات نظیر روش‌های شناسایی

به‌صورت بلادرنگ و یا روش‌های کشف تقلب انسانی را نیز مورد بررسی و پژوهش قرار داد.

۷- مراجع 7- References

رجب‌نیا جواد، ذبیحی مهدیه، وفایی‌جهان مجید، "تشخیص روبات‌های وب با استفاده از سیستم استنتاج فازی مبتنی بر درخت تصمیم"، *هفتمین کنفرانس داده‌کاوی ایران*، ۱۳۹۲.

[1] J. Rajab Nia, M. Zabihi, M. VafahiJahan, "web robot detection with fuzzy inference system based on decision trees," *The Seventh Iran Data Mining Conference*, 2013.

[2] B. W.N.Lo, R.. SharmaSedhain, "How Reliable Are Website Rankings? Implications For E-Business Advertising And Internet Search," *Issues in Information Systems*, Volume VII, No. 2, pp. 233–238, 2006.

[3] What is fake traffic?, [Online], https://sedo-us1.custhelp.com/app/answers/detail/a_id/678/~/what-is-fake-traffic, February 2017.

[4] D.S. Sisodia, Sh. Verma, O.P. Vyas, "A Comparative Analysis of Browsing Behavior of Human Visitors and Automatic Software Agents," *American Journal of Systems and Software*, vol. 3, no. 2, pp. 31-35, 2015.

[5] A. Stassopoulou, M.D. Dikaiakos," Web robot detection: A probabilistic reasoning approach," *Computer Networks*, Vol. 53, pp. 265–278, 2009.

[6] D. Doran, S.S. Gokhale, "Web Robot Detection Techniques: Overview And Limitations," *springer Data Mining and Knowledge Discovery*, Vol. 22, pp. 183-210, 2010.

[7] P.N. TAN, V. KUMAR,"Discovery of Web Robot Sessions Based on their Navigational Patterns," *Data Mining and Knowledge Discovery*, vol. 6, pp. 9-35, 2002.

[8] CH. Bomhardt, W. Gaul, L. Schmidt-Thieme, "Web Robot Detection - Preprocessing Web Logfiles for Robot Detection," *In Proceedings of SISCLADAG.Bologna*, Ital, pp. 113-124, 2005.

[9] D. Stevanovic, A. An, N. Vlajic, "Feature evaluation for web crawler detection with data mining techniques," *Elsevier, Expert Systems with Applications*, Vol. 39, pp. 8707–8717, 2012.

[10] D. Stevanovic, N. Vlajic, A. An, "Detection of malicious and non-malicious website visitors using unsupervised neural network learning," *Elsevier, Applied Soft Computing 13*, pp. 698–708, 2012.

[11] M. Zabihimayvan, M. VafaeiJahan, J. Hamidzadeh,"A Density Based Clustering Approach for Web Robot Detection," *IEEE, 4th International Conference On Computer And Knowledge Engineering (ICCKE)*, pp. 23–28, 2014.

[12] D.S. Sisodia, Sh. Verma, .O.P. Vyas, "Agglomerative Approach for Identification and Elimination of Web Robots from Web Server Logs to Extract Knowledge about Actual Visitors," *Journal of Data Analysis and Information Processing*, Vol. 3, pp. 1-10, 2015.

[13] J. Hamidzadeh, M. Zabihimayvan, R. Sadeghi, "Detection of Web site visitors based on fuzzy rough sets," *Springer*, pp. 2175–2188, 2017.

[14] user-agent-string. [online], http://user-agent-string.info/list-of-ua/bots-ip , ,(December 2017)
Bot vs.Browsers. [Online], http://www.botsvs-browsers.com, December 2017.

[15] User-Agents. [Online], http://www.user-agents.org, December 2017.

[16] S.S. Aksenova, "Machine Learning with WEKA :WEKA Explorer Tutorial for WEKA Version 3.4.3," 2004 .

[17]http://www.secrepo.com/maccdc2012/http.log.gz

[18] http://www.cs.waikato.ac.nz/ml/weka/

زهرا عبدی تحصیلات خود را در کارشناسی مهندسی کامپیوتر_نرم‌افزار در سال ۱۳۹۱ در دانشکدهٔ فنی دانشگاه مازندران به اتمام رساند و مدرک کارشناسی ارشد مهندسی فناوری اطلاعات_تجارت الکترونیک را از دانشگاه آزاد اسلامی واحد علوم و تحقیقات در سال ۱۳۹۷ دریافت کرد. موضوع پایان‌نامهٔ ایشان، ارائه راه‌کار برای شناسایی و مقابله با فریب ایجادشده به‌وسیلهٔ ربات‌ها به‌منظور بهبود رتبه‌بندی ترافیکی تارنماها بوده است.

نشانی رایانامهٔ ایشان عبارت است از:

zahraabdi.ce@gmail.com

مجتبی مازوچی تحصیلات خود را در مقطع کارشناسی مهندسی برق-مخابرات در دانشکده فنی دانشگاه تهران در سال ۱۳۷۱ به اتمام رساند و مدارک کارشناسی ارشد و دکترای مهندسی برق-مخابرات سیستم را به‌ترتیب از دانشگاه خواجه نصیرالدین طوسی و دانشگاه آزاد واحد علوم و تحقیقات تهران در سال‌های ۱۳۷۴ و ۱۳۹۳ دریافت کرد. از سال ۱۳۸۰ تاکنون عضو هیئت علمی پژوهشگاه ارتباطات و فناوری اطلاعات است و زمینه‌های پژوهشی مورد علاقه

ایشان مدیریت شبکه و کیفیت سرویس، شبکه‌های ارتباطی، مدیریت و به اشتراک‌گذاری طیف فرکانسی و تحلیل ترافیکی و ذائقه‌سنجی کاربران فضای مجازی است.
نشانی رایانامهٔ ایشان عبارت است از:
mazoochi@itrc.ac.ir

محمدعلی پورمینا دانشیار مهندسی برق (مخابرات) در دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران و استاد نمونه این دانشگاه در سال ۱۳۹۵ است. او مدرک دکترای مهندسی برق-مخابرات را از دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران در سال ۱۳۷۵ دریافت کرده و از همان سال به این دانشگاه پیوسته است. وی از سال ۱۳۷۱عضو مرکز تحقیقات مخابرات ایران بوده‌اند. از سال ۱۳۷۰ در حوزه شبکه‌های رادیوئی مبتنی بر بسته و سیستم‌های پردازش سیگنال دیجیتال پژوهش کرده و زمینه‌های پژوهشی فعلی مورد علاقه ایشان، سامانه‌های طیف گسترده، مخابرات موبایل سلولی، مخابرات بی‌سیم، پردازش‌گرهای DSP و شبکه‌های چندرسانه‌ای بی‌سیم است.
نشانی رایانامهٔ ایشان عبارت است از:
pourmina@srbiau.ac.ir