تشخیص بازدیدکنندگان مخرب و غیر مخرب وب سایت‌ها مبتنی بر روش ترکیبی فازی-عصبی

سوده لایقی
دانشجوی کارشناسی ارشد
مهندسی کامپیوتر- نرم افزار
دانشگاه آزاد اسلامی واحد مشهد
so_layeghi@yahoo.com

امیر حسین زارعی
دانشجوی کارشناسی ارشد
مهندسی کامپیوتر- نرم افزار
دانشگاه آزاد اسلامی واحد مشهد
zareie@ymail.com

مجید وفایی جهان
استادیار گروه کامپیوتر- نرم افزار
دانشگاه آزاد اسلامی واحد مشهد
VafaeiJahan@mshdiau.ac.ir

مهرداد جلالی
استادیار گروه کامپیوتر- نرم افزار
دانشگاه آزاد اسلامی واحد مشهد
jalali@mshdiau.ac.ir

چکیده: روبات‌های وب برنامه های نرم افزاری هستند که دائماً به صورت خودکار ساختار لینک‌های وب سایت‌ها را مورد پیمایش قرار می‌دهند. موتورهای جستجو، اسپم سازهای کامنت گذار و روبات‌های کپی کننده مطالب از انواع روبات‌های وب هستند. هدف روبات‌های وب کشف و بازیابی محتوا و دانش از وب می‌باشد. این روبات‌ها هم به منظور اعمال مفیدی مانند کشف لینک‌های خراب و هم اعمال مخربی مانند حمله توزیع شده مختل کننده سرویس (DDOS) طراحی شده‌اند. در این مقاله با استفاده از روش ترکیبی فازی-عصبی به تحلیل فایل ثبت وقایع به منظور تشخیص روبات‌های وب پرداخته شده است. روش پیشنهادی با روش‌های شبکه عصبی، ماشین بردار پشتیبان و C4.5 استیوانوویچ و شبکه بیزین استسپولو مقایسه شده است، که استفاده از سیستم فازی-عصبی و استخراج ویژگی‌های جدید باعث گردیده است که روش پیشنهادی دقت بالاتری نسبت به سایر روش‌ها در تشخیص روبات‌های وب داشته باشد.

واژه های کلیدی: روبات‌های وب، سیستم فازی-عصبی، فایل ثبت وقایع، کاربرد کاوی وب.

۱ – مقدمه

با گسترش شبکه جهانی اینترنت، بسیاری از زوایای زندگی انسان نیز تحت تأثیر این پدیده قرار گرفته است. به طوری که در کشورهای صنعتی، بسیاری از امور روزمره، از خریدهای روزانه گرفته تا آموزش و تجارت، همگی از طریق اینترنت صورت می‌گیرد. با پیشرفت تکنولوژی‌های مرتبط با کامپیوتر و افزایش قدرت برنامه‌نویسان هر روزه برنامه های سودمندی بر روی اینترنت به دنیا عرضه می‌شود. در مقابل، قدرت هکرها نیز افزایش یافته و برنامه های مخرب قدرتمندتری تولید شده است. پس نیاز است تا در دنیای مجازی بتوان تنها با یک برنامه خودکار نرم افزاری تفاوت میان یک کاربر انسان و یک برنامه نرم افزاری را تشخیص داد. روش‌های مختلفی برای این که بتوان یک کاربر انسان را از یک نرم افزار متمایز کرد وجود دارد.

تا کنون روش‌های مختلفی برای تشخیص روبات‌های وب[1] پیشنهاد شده است. بومهارت[2] و همکارانش روبات‌های وب را به چهار دسته: روش‌های ساده، تله، ارزیابی رفتار حرکتی روبات‌ها، مدل سازی الگوی رفتاری روبات‌های وب خلاصه کرده‌اند [1]. دسته بندی دوران[3] و همکارانش چهار دسته: تحلیل نحوی ثبت وقایع[4]، الگوی ترافیک[5]، تکنیک‌های یادگیری تحلیلی[6]، سیستم تست تورینگ[7] می‌باشد [2]. استیوانوویچ[8] و همکارانش بازدیدکنندگان

[1] Web robot detection
[2] Bomhardt
[3] Doran
[4] Syntactical log analysis
[5] Traffic pattern analysis
[6] Analytical learning techniques
[7] Turing test systems
[8] Stevanovic

وب سایت‌ها را چهار دسته: انسان، روبات‌ها با رفتار خوب، روبات‌ها با رفتار مخرب، بازدیدکننده ناشناخته در نظر گرفته‌اند و با روش‌های دسته بندی، مشخص کرده‌اند هر نشست مربوط به کدام دسته می‌باشد [3]. در این مقاله هدف اصلی، کشف دانش از فایل‌های ثبت وقایع[9] به منظور دسته بندی و تشخیص روبات‌های وب به کمک سیستم فازی-عصبی می‌باشد. این فرایند شامل سه فاز اصلی به شرح زیر می‌باشد. ۱- پیش پردازش[10] : که در این مرحله ورودی فایل ثبت وقایع و خروجی نشست کاربران می‌باشد. فاز پیش پردازش شامل یکسری ریز مرحله می‌باشد : (پاک‌سازی اطلاعات، شناسایی کاربران، شناسایی نشست). ۲-کشف الگو: که در این فاز ورودی نشست کاربران می‌باشد و برای دسته بندی از روش ترکیبی فازی-عصبی استفاده شده است. برای ساخت سیستم فازی-عصبی باید پارامتر های ورودی تعیین شود سپس آموزش سیستم فازی-عصبی با استفاده از الگوهای ورودی و بدست آوردن خروجی با توجه به پارامترهای تالی و در پایان ارزیابی آموزش سیستم فازی-عصبی با داده های آموزش. ۳- تحلیل الگوی کشف شده: در این فاز نرخ خطا و دقت دسته بندی بدست می‌آید.

بخش‌های مختلف این مقاله به این ترتیب می‌باشد: در بخش دوم کارهای مرتبط، بخش سوم آماده سازی مجموعه داده، بخش چهارم سیستم فازی-عصبی، بخش پنجم نتایج و شبیه سازی‌ها و در بخش ششم نتیجه گیری مقاله آورده شده است.

۲ – کارهای مرتبط

طبق دسته بندی دوران روش‌های تحلیل نحوی ثبت وقایع شامل: بررسی رشته های عامل کاربر[11] [4]، تکنیک تحلیل چند گامی[12] ثبت وقایع [5] می‌باشد و روش‌های تحلیل الگوی ترافیک شامل: تشخیص روبات‌های وب از طریق تحلیل نحوی و تحلیل الگو [6]، تشخیص روبات‌های وب بر اساس الگوهای منبع درخواست [7]، تشخیص بر اساس الگو های نرخ درخواست[13] [8]، تشخیص با استفاده از معیار ترافیک [9] می‌باشد. تکنیک‌های یادگیری تحلیلی شامل: تشخیص با استفاده از درخت تصمیم [10]، تشخیص با استفاده از شبکه عصبی [11]، تشخیص بر اساس شبکه بیزین [12,13]، تشخیص با استفاده از مدل مخفی مارکوف [14]می‌باشد و تکنیک‌های سیستم تست تورینگ شامل: تشخیص بر اساس تست کپچا[14] [15,16]، تشخیص با رفتار مروری انسان [17] می‌باشد.

۳ – آماده سازی مجموعه داده[15]

کلیه مراحل آماده سازی مجموعه دادها به این شرح می‌باشد:۱-ورودی فایل ثبت وقایع ۲- شناسایی نشست ۳- استخراج ویژگی برای هر نشست (استفاده از ویژگی‌های روش‌های قبلی و استخراج سه ویژگی جدید) ۴- بر چسب گذاری هر نشست (استفاده از بر چسب گذاری روش‌های قبلی و یک بر چسب گذاری جدید) ۵-مجموعه داده را به دو دسته مجموعه داده آموزش و تست تقسیم می‌کنیم ۶- استفاده از روش سیستم فازی-عصبی برای دسته بندی. با فرض صحت فرایند برچسب زدن، هدف اصلی بررسی دقت دسته بندی است [3].

۳-۱- ثبت وقایع

یک نمونه درخواست در فایل ثبت وقایع سرور پارس هاستینگ به شرح زیر است:

2012-04-13 06:11:44 GET /Site/MAGHALAT/mga86020404_files /image018.jpg 180.94.90.16 Mozilla/5.0+(compatible ;+MSIE+ 9.0;+Windows+NT+6.1;+Trident/5.0) - http://www.google.com .af /imgres 200 848

هر ورودی فایل ثبت وقایع به ترتیب حاوی اطلاعات زیر می‌باشد: تاریخ، ساعت، متد (GET,HEAD,…)، فایل درخواست شده، آدرس IP کلاینت، رشته عامل کاربر، کوکی، رشته ارجاع، کد پاسخ، تعداد بایتی که از کلاینت به سرور ارسال می‌شود.

۳-۲- شناسایی نشست

ابتدا تمام درخواست‌های HTTP بر اساس IP و User-agent یکسان گروه بندی می‌شوند سپس از یک رویکرد وقفه[16] برای شکستن این گروه ها به زیر گروه های دیگر استفاده می‌شود (اگر زمان وقفه[17] بین دو درخواست متوالی از یک زیرگروه IP بیش از یک حد آستانه باشد این طور فرض شود که آن

[9] Access log file
[10] Preprocessing
[11] User agent
[12] Multi-step
[13] Query rate
[14] CAPTCHA tests
[15] Dataset preparation

کاربر ، یک نشست جدید را شروع کرده است). معمولاً حد آستانه را ۳۰ دقیقه در نظر می گیرند. بدون شک یک عدم قطعیت در این رویکرد وجود دارد[3].

۳-۳- استخراج ویژگی برای هر نشست

پایه انتخاب ویژگی‌ها را بر اساس مطالعاتی که از رفتار روبات‌های وب در [1] [3] [10] [19-18] داشتیم در نظر می‌گیریم. ویژگی‌هایی که از هر نشست استخراج می‌شود به شرح زیر است: ۱-حداکثر نرخ کلیک[۱۸] ، ۲-مدت نشست[۱۹]، ۳-درصد درخواست تصویر[۲۰]، ۴-درصد درخواست صفحات HTML، ۵-درصد پاسخ خطای 4xx [۲۱]، ۶-درخواست فایل Robots.txt [۲۲]، ۷-درصد درخواست‌های با ارجاع خالی[۲۳]، ۸-درصدی از درخواست HTTP از نوع HEAD[۲۴]، ۹-نرخ درخواست دنباله متوالی[۲۵]، ۱۰-عمق درخواست صفحه[۲۶]، ۱۱-تعداد بایتی که از کلاینت به سرو ارسال می‌شود[۲۷]. ویژگی‌های یک تا یازده قبلاً برای تشخیص روبات‌های وب استفاده شده است بعلاوه سه ویژگی جدید که به شرح زیر آورده شده است، ویژگی‌های هستند که از هر نشست استخراج می‌شود.

درصدی از درخواست فایل CSS: مرورگرهای وب به صورت خودکار یک درخواست برای فایل CSS ارسال می‌کنند در حالی که روبات‌های وب نیازی به مشاهده فایل CSS ندارد، پس اگر در یک نشست تمام درخواست‌ها از یک نوع منبع باشد و فایل CSS مشاهده نشده باشد آنگاه آن نشست مشکوک به روبات وب می‌باشد.

درصدی از درخواست فایل‌های دیگر[۲۸] : در یک نشست اگر تنها یک نوع از منابع درخواست شده باشد به عنوان روبات در نظر گرفته می‌شود. برای تعداد منابعی که بازدید می‌شود باید یک حد آستانه در نظر گرفت اگر از یک حد آستانه بیشتر باشد آن نشست می‌تواند متعلق به کاربر انسان باشد، و همچنین اگر یک صفحه وب دیده شده اما نه همه منابع جاسازی شده در آن، آنگاه آن نشست می‌توان مشکوک به روبات وب باشد.

درصدی از کوکی‌ها[۲۹] : کوکی‌ها اطلاعاتی هستند که سرور HTTP می‌تواند به همراه منبع درخواست شده به ماشین کاربر ارسال کند. مرورگر کاربر ممکن است این اطلاعات را ذخیره کند و متعاقباً هنگام ارسال درخواست‌های بعدی اطلاعات آن را به سرور HTTPپس بفرستد. اگر درصد کوکی‌ها در یک نشست صفر باشد، آنگاه آن نشست می‌توان مشکوک به روبات باشد.

دلیل استفاده از سه ویژگی جدید این است که کاربر انسان برای مشاهده صفحات وب نیاز به مرورگرهای وب دارد در حالی که روبات‌های وب نیازی به استفاده از مرورگرهای وب ندارد به همین دلیل است: ۱- روبات‌های وب نیازی به مشاهده فایل CSS ندارند. ۲- روبات‌های وب نیازی به مشاهده کلیه منابع جاسازی شده در یک صفحه وب را ندارند. ۳- درصد کوکی‌ها در یک نشست روبات وب صفر می‌باشد.

۳-۴- بر چسب گذاری هر نشست[۳۰]

در ابتدا همه نشست‌ها به صورت پیش فرض انسان در نظر گرفته شده است به جز برچسب گذاری بعضی از نشست‌ها به عنوان روبات که به شرح زیر است : ۱- در صورتی که آدرس IP با لیست به روز شده IPهای روبات‌های وب شناخته شده یکی باشد، ۲- رشته عامل کاربر با لیست به روز شده عامل‌های کاربر روبات‌های وب شناخته شده یکی باشد[24-20]، ۳- درخواست برای فایل ROBOTS.TXT داده شده باشد، ۴- مدت نشست بیشتر از

---

[16] Timeout
[17] Time-lapse
[18] Maximum click rate
[19] Duration of session
[20] Percentage of image requests
[21] Percentage of 4xx error responses
[22] Robots.txt file request
[23] Percentage of requests with unassigned referrers
[24] Percentage of HTTP requests of type HEAD
[25] consecutive sequential request ratio
[26] page request depth
[27] cs_Bytes
[28] other file
[29] Cookies
[30] Labeling training data

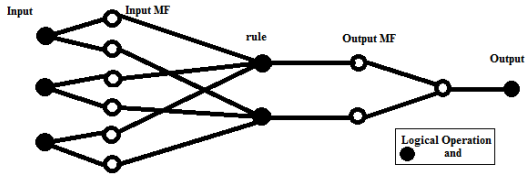
سه ساعت، ۵-نسبت درخواست HTML به تصویر[۳۱] در یک نشست بزرگ‌تر از ۱۰ باشد [13]. برچسب گذاری یک تا پنج قبلاً برای تشخیص روبات‌های وب استفاده شده است، بعلاوه یک برچسب گذاری جدید که به شرح زیر آورده شده است، مواردی است که در مرحله برچسب گذاری استفاده شده است.
در صورتی که در یک نشست تمام درخواست‌ها از یک نوع منبع باشد و فایل CSS مشاهده نشده باشد و درصد کوکی‌ها در یک نشست صفر باشد آنگاه به عنوان روبات وب برچسب گذاری می شود. سپس بازرسی به صورت دستی توسط یک فرد خبره برای تایید یا رد برچسب گذاری که پیشنهاد شده است انجام شده است. هدف روش نیمه اتوماتیک به حداقل رساندن نویز در مجموعه آموزش تعرف شده است.

۴- روش ترکیبی فازی-عصبی[۳۲]

روش استفاده شده در این مقاله ترکیبی از منطق فازی و سیستم عصبی می‌باشد، بدین صورت علاوه بر استنتاج فازی، سادگی و قابلیت تفسیر را نیز به خواص شبکه عصبی اضافه می‌کند که باعث عملکرد بهتر نسبت به روش‌های پیشین می‌گردد. این روش برای بهبود بخشیدن به قوانین سیستم فازی با کمک الگوریتم آموزش داده های تست در شبکه عصبی می‌باشد. سیستم فازی-عصبی در مقایسه با شبکه عصبی به خاطر تنظیم پذیری پارامترهای سیستم فازی، سریع‌تر آموزش می‌بیند و همچنین دقت بالاتری دارد.

سیستم فازی- عصبی را می‌توان به دو شکل پیاده سازی نمود. در شکل اول می‌توان از عملیات فازی و یا اعداد فازی در شبکه های عصبی استفاده نمود. در شکل دوم که در این مقاله از آن استفاده شده است می‌توان سیستم فازی را به صورت عصبی تحقق بخشید [28-27]. به عبارت دیگر می‌توان با استفاده از اطلاعات ورودی و خروجی، سیستم فازی را به وسیله شبکه عصبی به نحوی آموزش داد که توابع عضویت[۳۳] ورودی، توابع درجه یک خروجی (در شبکه فازی Takagi-Sugeno-kang) و قوانین فازی خود را با این اطلاعات سازگار کنند.

در تشخیص روبات‌های وب با روش ترکیبی فازی-عصبی ورودی‌ها به این سیستم ۱۴ ویژگی استخراج شده می‌باشد. نحوه آموزش سیستم به این صورت می‌باشد که سیستم فازی-عصبی از داده های آموزش و ۱۴ ویژگی مربوط به هر نشست برای آموزش استفاده کرده است. در شکل ۱ آرایش شبکه عصبی-فازی با دو تابع عضویت ورودی و سه ویژگی را نشان می‌دهد.



شکل ۱- آرایش ANFIS با دو تابع عضویت ورودی و سه ویژگی

در لایه دوم که لایه فازی کننده نیز نامیده می‌شود میزان تعلق ورودی‌ها به مجموعه های فازی ورودی محاسبه می‌شود. لایه سوم لایه تصمیم گیری و استنتاج فازی است، در این لایه با توجه به تعداد مجموعه های فازی قانون وجود دارد. در لایه چهارم یا لایه فازی زدایی، قسمت تالی قوانین کنترلی قرار داشته و عمل فازی زدایی نیز به صورت توأم در آن انجام می‌شود. در این لایه متناظر با هر خروجی کنترل کننده، یک نورون قرار دارد. خروجی شبکه نیز احتمال روبات بودن هر نشست می‌باشد.

در شبکه های عصبی- فازی از قوانین اگر-آنگاه فازی Takagi-sugeno مرتبه یک استفاده شده است خروجی هر قانون ترکیب خطی متغیرهای ورودی و مقادیر ثابت می‌باشد و خروجی نهایی میانگین وزنی هر یک از خروجی‌های قوانین می‌باشد. در نرم افزار MATLAB دو مرحله برای طراحی ANFIS وجود دارد. مرحله اول، طراحی پارامترهای مقدم و مرحله بعد آموزش پارامترهای تالی می‌باشد. در اینجا برای طراحی پارامترهای مقدم (تعیین شکل توابع عضویت) از روش Grid Partition استفاده شده است. زمانی که پارامترهای مقدم به دست آمدند، پارامترهای تالی بر اساس داده های ورودی-خروجی، به دست می‌آیند. برای آموزش ANFIS، نتایج شبیه سازی‌ها نشان داد که روش آموزشی Back Propagtion یا Hybrid به تنهایی مناسب نیست. لذا ایده ای که در این مقاله از آن استفاده شد به این صورت است که ابتدا سیستم با روش Hybrid آموزش داده شود، سپس سیستم آموزش دیده به وسیله روش Back Propagtion مجدد آموزش داده می‌شود. در مجموعه داده آموزش و تست اینکه هر نشست متعلق به روبات وب یا کاربر انسان است توسط برچسب گذاری مشخص شده است. جهت آموزش شبکه فازی-عصبی ابتدا از ویژگی‌های مجموعه داده آموزش استفاده شده است سپس برای اعتبار سنجی شبکه ANFIS، احتمال روبات وب بودن نشست‌های باقی مانده (مجموعه داده تست) محاسبه شده است.

[31] HTML-to-image request ratio
[32] ANFIS
[33] Membership Function

۵- نتایج و شبیه سازی‌ها

جهت بررسی سیستم پیشنهادی از فایل ثبت وقایع سرور پارس هاستینگ و پارس وب سایت استفاده شده است [26-25]. نشست‌های استخراج شده از فایل ثبت وقایع شامل انواع روبات‌های وب متنی و غیر متنی از قبیل موتورهای جستجوی متفاوت، جمع کنندگان تصاویر و ... می‌باشد. ۸۰ درصد از مجموعه داده به عنوان مجموعه داده های آموزشی و ۲۰ درصد باقی مانده را به عنوان مجموعه داده های تست در نظر گرفته‌ایم. مجموعه داده آموزش پارس هاستینگ [25] سه گروه به قرار زیر می‌باشد : ۱-مجموعه داده آموزش را هیچ تغییری نداده است (PH_D1). ۲- در مجموعه داده آموزش نرخ روبات وب را به ۵۰ ٪ رسانده است (PH_D2). ۳- در مجموعه داده آموزش نرخ کاربر انسان را به ۵۰ ٪ رسانده است (PH_D3). و همچنین برای مجموعه داده آموزش پارس وب سایت [26] سه گروه به قرار زیر می‌باشد : ۱- مجموعه داده آموزش را هیچ تغییری نداده است (PW_D1). ۲- در مجموعه داده آموزش نرخ کاربر انسان را به ۵۰ ٪ رسانده است (PW_D2). ۳- در مجموعه داده آموزش نرخ روبات وب را به ۵۰ ٪ رسانده است (PW_D3). در جدول ۱ مشخصات دو مجموعه داده پارس هاستینگ و پارس وب سایت نشان داده شده است، همان طور که در این جدول مشاهده می‌کنید تعداد نشست‌های استخراج شده در مجموعه داده پارس وب سایت بیشتر از مجموعه داده پارس هاستینگ است. برای ارزیابی دقیق‌تر کارایی روش پیشنهادی آن را با روش استسوپلو[34] مبتنی بر مدل کلاس‌بندی شبکه بیزین [13] و با روش‌های شبکه عصبی، ماشین بردار پشتیبان و C4.5 استیوانوویچ [3] که دو ویژگی جدید نسبت به روش استسوپلو ارائه کرده است مقایسه شده است. در جدول ۲ نرخ خطای دسته بندی برای مجموعه داده های مختلف آورده شده است، همان‌طور که در این جدول مشاهده می‌کنید هر چه تعداد مجموعه داده آموزش بیشتر باشد نرخ خطای دسته بندی کمتر خواهد بود. در شکل ۲ نمودار میله ای مقایسه نرخ خطای دسته بندی میانگین الگوریتم‌ها را می‌توانید مشاهده کنید.

جدول ۱- مشخصات مجموعه داده ها

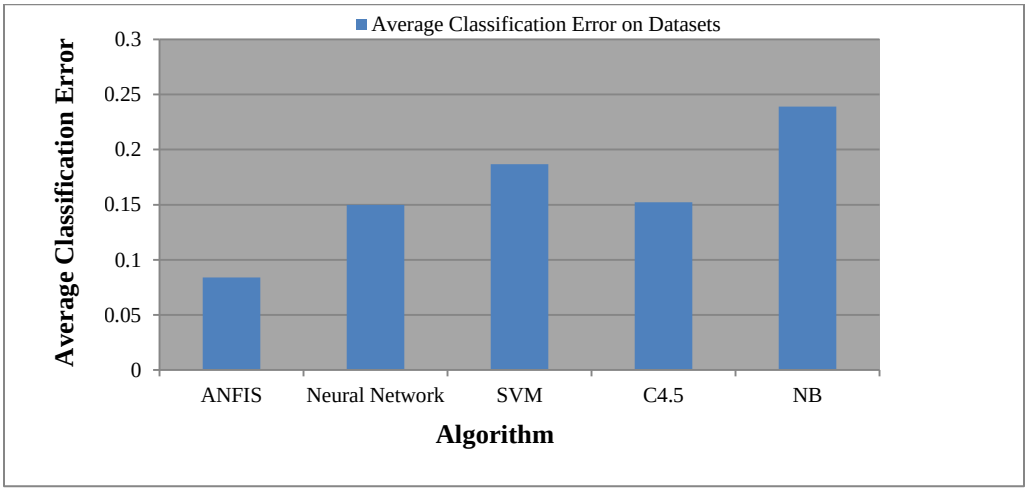
مشخصات / مجموعه داده	تعداد ویژگی‌ها	تعداد کل نشست‌ها	مجموعه آموزش	تعداد کاربر انسان در مجموعه آموزش	تعداد روبات وب در مجموعه آموزش	مجموعه تست
PH_D1	۱۴	۱۱۷۷	۹۴۲	۵۴۹	۳۹۳	۲۳۵
PH_D2	۱۴	۱۳۳۳	۱۰۹۸	۵۴۹	۵۴۹	۲۳۵
PH_D3	۱۴	۱۰۲۱	۷۸۶	۳۹۳	۳۹۳	۲۳۵
PW_D1	۱۴	۲۴۲۲	۱۹۳۸	۷۵۷	۱۱۸۱	۴۸۴
PW_D2	۱۴	۲۸۴۶	۲۳۶۲	۱۱۸۱	۱۱۸۱	۴۸۴
PW_D3	۱۴	۱۹۹۸	۱۵۱۴	۷۵۷	۷۵۷	۴۸۴

جدول ۲- مقایسه نرخ خطای دسته بندی برای مجموعه داده های مختلف

الگوریتم / مجموعه داده	ANFIS	Neural Network	SVM	C4.5	NB
PH_D1	۰,۱۴۳۱	۰,۲۴۵۸	۰,۲۵۳۴	۰,۲۱۶۴	۰,۳۰۸۵
PH_D2	۰,۰۵۳۳	۰,۱۴۸۹	۰,۱۸۴۷	۰,۱۵۹۳	۰,۲۳۰۶
PH_D3	۰,۱۱۲۳	۰,۱۵۰۵	۰,۱۹۳۸	۰,۱۶۰۳	۰,۲۹۲۲

[34] Stassopoulou

PW_D1	۰,۰۴۴۹	۰,۱۲۶۶	۰,۱۷۵۱	۰,۱۲۷۳	۰,۱۸۲۳
PW_D2	۰,۰۲۴۵	۰,۰۴۸۹	۰,۰۷۶	۰,۰۶۵۱	۰,۱۶۰۳
PW_D3	۰,۱۳۶۷	۰,۱۷۹۱	۰,۲۳۸۴	۰,۱۸۴۷	۰,۲۶۰۹
میانگین	۰,۰۸۴۱	۰,۱۴۹۹	۰,۱۸۶۹	۰,۱۵۲۱	۰,۲۳۹۱



شکل ۲- نمودار میله ای مقایسه نرخ خطای دسته بندی میانگین الگوریتم‌ها

۶- نتیجه گیری

در این مقاله روشی مبتنی بر شبکه فازی-عصبی برای تشخیص روبات‌های وب با استفاده از فایل ثبت وقایع مربوط بـه سـرورهای پـارس هاسـتینگ[۳۵] و پارس وب سایت ارائه شده است؛ و همچنین با به‌کارگیری سه ویژگی و یک برچسب گذاری جدید در سیستم پیشنهادی، تشخیص روبات‌های وب از دقت بالایی برخوردار شده است.

تشخیص و دسته بندی روبات‌های وبی که تلاش در تقلید رفتار انسان دارند به عنوان مهم‌ترین چالش دسته بندی است، در واقع روبات‌های وبی کـه به اشتباه در دسته انسان قرار می‌گیرند. با به‌کارگیری روش پیشنهادی تا حدودی این چالش کاهش داده شده است. در این مقاله با استفاده از نرخ خطای دسته بندی سیستم پیشنهادی مورد ارزیابی قرار گرفته است روش پیشنهادی با روش استسوپلو مبتنی بر مدل کلاس‌بندی شبکه بیـزین و بـا روش‌هـای شبکه عصبی، ماشین بردار پشتیبان و C4.5 استیوانوویچ که دو ویژگی جدید نسبت به روش استسوپلو ارائه کرده است مقایسه شده اسـت. نتـایج بدسـت آمده از ارزیابی‌های مختلف نشان داد که دقت تشخیص روبات‌های وب در مدل پیشنهادی بهتر از سایر روش‌ها است و همچنین مشاهده شد که هـر چـه تعداد مجموعه داده آموزش بیشتر باشد نرخ خطای دسته بندی کمتر خواهد بود.

مراجع

[1] Bomhardt, C., Gaul, W., Schmidt-Thieme, L., "Web Robot detection preprocessing web logfiles for Robot Detection", New Developments in Classification and Data Analysis Studies in Classification, Data Analysis, and Knowledge Organization, pp.113-124, 2005.

[2] Doran, D., Gokhale, Swapna S., "Web robot detection techniques: overview and limitations" Data Mining and Knowledge Discovery, Vol.22, pp.183-210, 2011.

[35] Pars Hosting

[3] Stevanovic, D., An, A., Vlajic, N., "Feature evaluation for web crawler detection with data mining techniques", Expert Systems with Applications: An International Journal, Vol.39, pp.8707-8717, 2012.
[4] Kabe, T., Miyazaki, M., "Determining WWW user-agents from server access log", In: Proceedings of seventh international conference on parallel and distributed systems, pp 173–178, 2000.
[5] Huntington, P., Nicholas, D., Jamali, Hamid R., "Web robot detection in the scholarly information environment", Journal of Information Science, Vol.34, pp.726-741, 2008.
[6] Geens, N., Huysmans, J., Vanthienen, J., "Evaluation of Web Robot Discovery Techniques: A Benchmarking Study", in Proc. Industrial Conference on Data Mining, pp.121-130, 2006.
[7] Guo, W., Ju, S., Gu, Y., "Web robot detection techniques based on statistics of their requested URL resources", in Proc. CSCWD (1), pp.302-306, 2005.
[8] Duskin, O., Feitelson, Dror G., "Distinguishing humans from robots in web search logs: preliminary results using query rates and intervals", In: Proceedings of 2009 workshop on Web Search Click Data, pp 15–19, 2009.
[9] Lin, X., Quan, L., Wu, H., "An Automatic Scheme to Categorize User Sessions in Modern HTTP Traffic", in Proc. GLOBECOM, pp.1485-1490, 2008.
[10] Tan, P., Kumar, V., "Discovery of Web Robot Sessions Based on their Navigational Patterns", Data Mining and Knowledge Discovery, Vol.6, pp.9-35, 2002.
[11] Stevanovic, D., Vlajic, N., An, A., "Detection of malicious and non-malicious website visitors using unsupervised neural network learning", Applied Soft Computing, Vol.13, pp.698-708, 2013.
[12] Stassopoulou, A., Dikaiakos, Marios D., "A Probabilistic Reasoning Approach for Discovering Web Crawler Sessions", in Proc. APWeb/WAIM, pp.265-272, 2007.
[13] Stassopoulou, A., Dikaiakos, Marios D., "Web robot detection: A probabilistic reasoning approach", Computer Networks: The International Journal of Computer and Telecommunications Networking,Vol.53, pp.265-278, 2009.
[14] Lu, WZ., Yu, SZ., "Web robot detection based on hidden Markov model", In: Proceedings of international conference on communications, circuits and systems, pp 1806–1810, 2006.
[15] Ahn, Luis V., Blum, M., Hopper, Nicholas J., Langford, J., "CAPTCHA: Using Hard AI Problems for Security", Proceedings of the 22nd international conference on Theory and applications of cryptographic techniques, pp.294-311, 2003.
[16] Motoyama, M., Levchenko, K., Kanich, C., McCoy, D., Voelker, Geoffrey M., Savage, S. "Re: CAPTCHAs-Understanding CAPTCHA-Solving Services in an Economic Context", Proceedings of the 19th USENIX conference on Security, pp.435-462, 2010.
[17] Park, K., Pai, Vivek S., Lee, K., Calo, Seraphin B., "Securing Web Service by Automatic Robot Detection", In Proceedings of USENIX Annual Technical Conference, General Track., pp.255-260, 2006.
[18] Dikaiakos, Marios D., Stassopoulou, A., Papageorgiou, L., "Characterizing Crawler Behavior from Web Server Access Logs", in Proc. EC-Web, pp.369-378, 2003.
[19] Dikaiakos, Marios D., Stassopoulou, A., Papageorgiou, L., "An investigation of WWW crawler behavior: characterization and metrics", Computer Communications, Vol.28, pp.880–897, 2005.
[20] User-Agents. [Online], August 2011, http://www.user-agents.org.
[21] Bot vs.Browsers. [Online], August 2011, http://www.botsvsbrowsers .com.
[22] User agent string. [Online], August 2011, http://www.useragentstring.com.
[23] Robotstxt. [Online], 2007, www.robotstxt.org.
[24] user-agent-string. [online], 2012, http://user-agent-string.info/list-of-ua/bots-ip.
[25] Pars Hosting. November 2012, http://parshosting.com.
[26] Pars Web Site. November 2012, http://www.parswebsite.com.
[27] Lin, C., Lu, Y., "A Neural Fuzzy system with Fuzzy Supervised Learning", IEEE, Trans. Syst. Man and Cyber, Vol. 26, No. 5,pp.744-763, 1996.
[28] Jang, J.R., "ANFIS: Adaptive-Neural-Based Fuzzy Inference System", IEEE, Trans, Syst. Man Cyber., Vol.23, No.3, pp.665-684, 1993.