ارزیابی ویژگی‌ها برای تشخیص بازدیدکنندگان مخرب و غیر مخرب وب سایت‌ها مبتنی بر روش‌های داده کاوی

سوده لایقی[۱]، امیرحسین زارعی[۲]، مجید وفایی جهان[۳] و مهرداد جلالی[۴]

۱- دانشگاه آزاد اسلامی مشهد گروه کامپیوتر نرم افزار، So_layeghi@ymail.com

۲- دانشگاه آزاد اسلامی مشهد گروه کامپیوتر نرم افزار، Zareie@yahoo.com

۳- دانشگاه آزاد اسلامی مشهد گروه کامپیوتر نرم افزار، Vafaeijahan@mshdiau.ac.ir

٤- دانشگاه آزاد اسلامی مشهد گروه کامپیوتر نرم افزار، Jalali@mshdiau.ac.ir

چکیده

در این مقاله بازدیدکنندگان وب سایت‌ها به چهار گروه انسان، روبات وب با رفتار خوب، روبات وب با رفتار مخرب و بازدیدکنندگان ناشناخته تقسیم بندی شده است. روبات‌های وب یا خزنده های وب برنامه های نرم افزاری هستند که دائماً به صورت خودکار ساختار لینک‌های وب سایت‌ها را مورد پیمایش قرار می‌دهند. 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 طراحی شده‌اند. تشخیص و دسته بندی روبات‌های وبی که تلاش در تقلید رفتار انسان دارند به عنوان مهم‌ترین چالش دسته بندی است. در این مقاله برای تشخیص بازدیدکنندگان مخرب و غیر مخرب وب سایت‌ها سه ویژگی جدید معرفی شده است. ویژگی‌های بیان شده در مقالات گذشته بعلاوه سه ویژگی جدید با استفاده از روش‌های شبکه عصبی، ماشین بردار پشتیبان، C4.5، شبکه بیزین و شبکه باور بیزی مقایسه شده است. استخراج ویژگی‌های جدید برای تشخیص بازدیدکنندگان وب سایت‌ها باعث شد که دقت دسته بندی در مقایسه با روش‌های دیگر با ویژگی‌های کمتر، بهبود یابد و همچنین نشان داده شده است هر چه تعداد مجموعه داده آموزش بیشتر باشد دقت دسته بندی بهتر خواهد بود.

واژه های کلیدی: روبات‌های وب، روش‌های داده کاوی، 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.

تا کنون روش‌های مختلفی برای تشخیص روبات‌های وب پیشنهاد شده است. بومهارت و همکارانش روبات‌های وب را به چهار دسته: روش‌های ساده، تله، ارزیابی رفتار حرکتی روبات‌ها، مدل سازی الگوی رفتاری روبات‌های وب خلاصه کرده‌اند (۱). دسته بندی دوران و همکارانش چهار دسته: تحلیل نحوی ثبت وقایع، الگوی ترافیک، تکنیک‌های یادگیری تحلیلی، سیستم تست تورینگ می‌باشد (۲). طبق دسته بندی دوران روش‌های تحلیل نحوی ثبت وقایع شامل: بررسی رشته های عامل کاربر (۴)، تکنیک تحلیل چند گامی ثبت وقایع (۵) می‌باشد و روش‌های تحلیل الگوی ترافیک شامل: تشخیص روبات‌های وب از طریق تحلیل نحوی و تحلیل الگو (۶)، تشخیص روبات‌های وب بر اساس الگوهای منبع درخواست (۷)، تشخیص بر اساس الگو های نرخ درخواست (۸)، تشخیص با استفاده از معیار ترافیک (۹) می‌باشد. تکنیک‌های یادگیری تحلیلی شامل: تشخیص با استفاده از درخت تصمیم (۱۰)، تشخیص با استفاده از شبکه عصبی (۱۱)، تشخیص بر اساس شبکه بیزین (۱۲و۱۳)، تشخیص با استفاده از مدل مخفی مارکوف (۱۴) می‌باشد و تکنیک‌های سیستم تست تورینگ شامل: تشخیص بر اساس تست کپچا (۱۵و۱۶)، تشخیص با رفتار مروری انسان (۱۷) می‌باشد. استیوانوویچ و همکارانش دو ویژگی جدید معرفی کردند و بازدیدکنندگان وب سایت‌ها را چهار دسته: انسان، روبات‌ها با رفتار خوب، روبات‌ها با رفتار مخرب، بازدیدکننده ناشناخته در نظر گرفته‌اند و با روش‌های داده کاوی، ویژگی‌های بیان شده در مقالات گذشته بعلاوه دو ویژگی‌های جدید را مورد ارزیابی قرار دادند (۳).

در این مقاله هدف اصلی، کشف دانش از فایل‌های ثبت وقایع به منظور دسته بندی و تشخیص بازدیدکنندگان وب مخرب و غیر مخرب به کمک روش‌های داده کاوی می‌باشد. این فرایند شامل سه فاز اصلی به شرح زیر می‌باشد. ۱- پیش پردازش : 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 روش‌های داده کاوی استفاده شده است. ۳- تحلیل الگوی کشف شده: در این فاز با استفاده از معیار دقت دسته بندی به تحلیل الگوی کشف شده پرداخته شده است.

ادامه این مقاله به صورت زیر سازماندهی شده است: در بخش دوم آماده سازی مجموعه داده، بخش سوم آزمایشات و نتایج و در بخش چهارم نتیجه گیری مقاله ارائه شده است.

۲- آماده سازی مجموعه داده

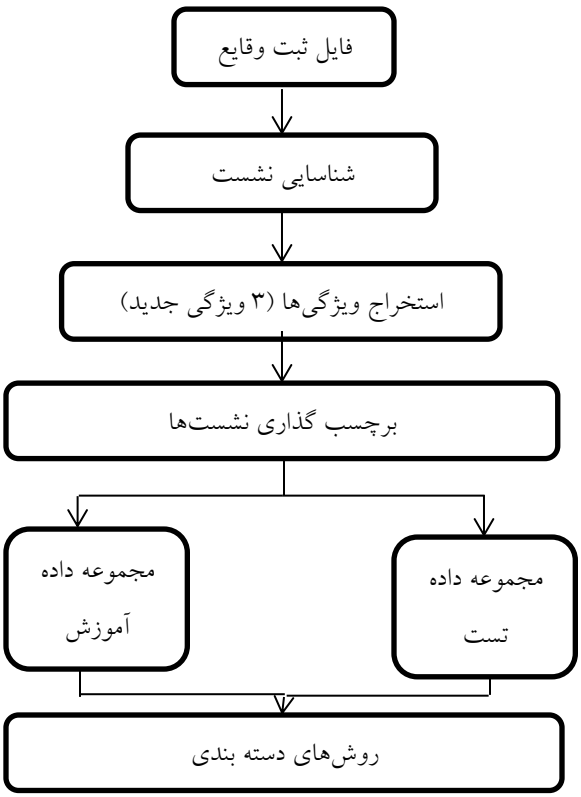
کلیه مراحل آماده سازی مجموعه دادها که در شکل ۱ مشاهده می‌شود به این شرح است: ۱- ورودی فایل ثبت وقایع ۲- شناسایی نشست ۳- استخراج ویژگی برای هر نشست (استفاده از ویژگی‌های روش‌های قبلی و استخراج سه ویژگی جدید) ۴- بر چسب گذاری هر نشست ۵-مجموعه داده را به دو دسته مجموعه داده آموزش و تست تقسیم می‌کنیم ۶- استفاده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

۲-۱- فایل ثبت وقایع

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.

۲-۲- شناسایی نشست

ابتدا تمام درخواست‌های HTTP بر اساس IP و User-agent یکسان گروه بندی می‌شوند سپس از یک رویکرد وقفه برای شکستن این گروه ها به زیر گروه های دیگر استفاده می‌شود (اگر زمان وقفه بین دو درخواست متوالی از یک زیرگروه IP بیش از یک حد آستانه باشد این طور فرض شود که آن کاربر، یک نشست جدید را شروع کرده است). معمولاً حد آستانه را ۳۰ دقیقه در نظر می‌گیرند. بدون شک یک عدم قطعیت در این رویکرد وجود دارد (۳).



شکل ۱- مراحل آماده سازی مجموعه داده

۲-۳- استخراج ویژگی برای هر نشست

پایه انتخاب ویژگی‌ها را بر اساس مطالعاتی که از رفتار روبات‌های وب در (۱) (۱۸و۱۹) (۱۰) (۳) داشتیم در نظر می‌گیریم. ویژگی‌های که از هر نشست استخراج می‌شود به شرح زیر است: ۱-حداکثر نرخ کلیک، ۲-نسبت درخواست HTML به تصویر، ۳-درصد درخواست فایل‌های PDF یا PS، ۴-درصد پاسخ خطای 4xx، ۵-درخواست فایل Robots.txt، ۶-درصد درخواست‌های با ارجاع خالی، ۷-درصدی از درخواست HTTPاز نوع HEAD، ۸-نرخ درخواست دنباله متوالی، ۹-عمق درخواست صفحه، ۱۰-تعداد بایتی که از کلاینت به سرو ارسال می‌شود، ۱۱-مدت نشست. ویژگی‌های یک تا یازده قبلاً برای تشخیص روبات‌های وب استفاده شده است بعلاوه سه ویژگی جدید که در ادامه آورده شده است، ویژگی‌های هستند که از هر نشست استخراج می‌شود. کاربر انسان برای مشاهده صفحات وب نیاز به مرورگرهای وب دارد در حالی که روبات‌های وب نیازی به استفاده از مرورگرهای وب ندارد. با توجه به این مورد این سه ویژگی جدید به این شرح می‌باشد.

۲-۳-۱- 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.

۲-۳-۲- درصدی از درخواست فایل‌های دیگر

مرورگرهای وب به صورت خودکار کلیه منابع جاسازی شده در یک صفحه وب را مشاهده می‌کنند در یک نشست اگر تنها یک نوع از منابع درخواست شده باشد آن نشست مشکوک به روبات وب می‌باشد. بنابراین اگر یک صفحه وب دیده شده اما نه همه منابع جاسازی شده در آن، آنگاه آن نشست می‌توان مشکوک به روبات وب باشد.

۲-۳-۳- درصدی از کوکی‌ها

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 HTTP پس بفرستد. اگر درصد کوکی‌ها در یک نشست صفر باشد، آنگاه آن نشست می‌توان مشکوک به روبات باش‌د.

در شکل ۲ تعدادی از نشست‌ها با چهارده ویژگی استخراج شده و یک ستون Visitor_4class مربوط به بر چسب گذاری با چهار دسته : بازدیدکننده ناشناخته برچسب صفر، انسان برچسب یک، روبات‌ها با رفتار غیر مخرب بر چسب دو، روبات‌ها با رفتار مخرب برچسب سه، از فایل ثبت وقایع سرور پارس هاستینگ نشان داده شده است.

۲-۴- بر چسب گذاری هر نشست

برچسب گذاری نشست‌ها بر اساس چهار کلاس به شرح زیر است (۲۰-۲۴) : ۱- مقایسه عامل کاربر با لیست به روز شده عامل های کاربر مرورگرهای شناخته شده: در صورتی که رشته عامل کاربر با لیست به روز شده عامل های کاربر

p_referer	max_click	d_session	p_jpg	p_htm	p_4XX	robots	p_m_h	C_S	mean	count_css	p_other_file	p_cookie	mean_cs_b	Visitor_4clas
0	0	10	83.33	0	0	0	0	91.67	1.833333	1	8.333333015	0	360.16666	1
14.28571	0	26	85.71	0	14.29	0	0	78.57	1.714286	1	7.142857075	0	348	1
0	0	3	83.33	0	0	0	0	58.33	1.833333	1	8.333333015	0	314.5	1
16.66667	0	4	83.33	0	8.333	0	0	75	1.75	1	8.333333015	0	345.58334	1
0	0	2	83.33	0	0	0	0	75	1.833333	1	8.333333015	0	358.75	1
100	0	14	0	87.5	12.5	1	0	100	1	0	0	0	308.75	3
100	0	85860	0	0	0	0	99.46236	100	1	0	0	0	190.99463	0
0	0	2	83.33	0	0	0	0	58.33	1.833333	1	8.333333015	0	392	1
100	0	16	0	87.5	12.5	1	0	100	1	0	0	0	352.75	3
14.28571	0	327	80.95	9.524	14.29	0	0	71.43	1.571429	1	4.761904716	0	396.28571	1
8.333333	0	7	83.33	0	8.333	0	0	50	1.75	1	8.333333015	0	380	1
100	0	85598	0	0	0	0	100	100	1	0	0	0	191	0
10.34483	0	70	82.76	6.897	6.897	0	0	58.62	1.689655	2	6.896551609	0	403.48276	1
11.11111	0	240	83.33	5.556	11.11	0	0	44.44	1.666667	1	0	0	393.16666	1
16.66667	0	4	83.33	0	8.333	0	0	66.67	1.75	1	8.333333015	0	368.91666	1
100	0	85938	0	0	0	0	100	100	1	0	0	0	191	0
7.692307	0	33	84.62	0	7.692	0	0	61.54	1.769231	1	7.692307472	0	472.46155	1
0	0	5	84.62	0	0	0	0	69.23	1.846154	1	7.692307472	0	348	1
8.333333	0	4	83.33	0	8.333	0	0	58.33	1.75	1	8.333333015	0	399.5	1
15	0	12	80	10	10	0	0	60	1.6	1	5	0	408.70001	1
100	2	4	0	100	100	0	0	12.5	1.875	0	0	0	69.25	0
100	0	85634	0	0	0	0	100	100	1	0	0	0	191	0
10.52632	0	75	84.21	5.263	5.263	0	0	57.89	1.631579	1	5.263157845	0	380.78946	1
100	0	4547	50	41.67	8.333	1	0	50	1.5	0	0	0	221.33333	3

شکل ۲: نمونه ای از داده پارس هاستینگ

مرورگرهای شناخته شده مطابقت داشته باشد و به فایل Robots.txt دسترسی پیدا نکرده باشد به عنوان کاربر انسان برچسب گذاری می‌شود. ۲- مقایسه عامل کاربر با لیست به روز شده عامل‌های کاربر روبات‌های وب با رفتار خوب شناخته شده: در صورتی که رشته عامل کاربر با لیست به روز شده عامل‌های کاربر روبات‌های وب با رفتار خوب شناخته شده مطابقت داشته باشد به عنوان روبات وب با رفتار خوب برچسب گذاری می‌شود. ۳- مقایسه عامل کاربر با لیست به روز شده عامل‌های کاربر روبات‌های وب مخرب شناخته شده: در صورتی که رشته عامل کاربر با لیست به روز شده عامل‌های کاربر روبات‌های وب مخرب شناخته شده مطابقت داشته باشد به عنوان روبات وب مخرب برچسب گذاری می‌شود. ۴- بقیه نشست‌ها به عنوان کاربر ناشناخته برچسب گذاری می‌شود.

۲-۵- مجموعه داده ها

در جدول ۱ مشخصات دو مجموعه داده پارس هاستینگ و پارس وب سایت نشان داده شده است (۲۵و۲۶). 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 و نتایج بدست آمده از مجموعه داده پارس وب سایت بر روی روش شبکه باور بیزی از دقت کلی بالاتری برخوردار می‌باشد. ۸۰ درصد از مجموعه داده به عنوان مجموعه داده های آموزشی و ۲۰ درصد باقی مانده به عنوان مجموعه داده های تست در نظر گرفته شده است که مشخصات این دو مجموعه داده در جدول ۲ و جدول ۳ قابل مشاهده است.

جدول ۱ : مشخصات مجموعه داده ها

مشخصات / مجموعه داده	تعداد کل نشست‌ها	تعداد بازدید کننده ناشناخته	تعداد کاربر انسان	تعداد روبات وب غیر مخرب	تعداد روبات وب مخرب
PH	۲۴۲۲	۱۰۳۹	۱۰۹۲	۱۲۷	۱۶۴
PW	۱۳۷۳۴	۳۵۶۶	۸۵۰۷	۱۴۵۹	۲۰۲

جدول ۲ : مشخصات مجموعه داده ها آموزش

مشخصات / مجموعه داده	تعداد کل نشست‌ها	تعداد بازدید کننده ناشناخته	تعداد کاربر انسان	تعداد روبات وب غیر مخرب	تعداد روبات وب مخرب
PH	۱۹۳۸	۹۵۲	۷۷۱	۱۰۲	۱۱۳
PW	۱۰۹۸۷	۳۲۴۶	۶۴۳۶	۱۱۴۴	۱۶۱

جدول ۳ : مشخصات مجموعه داده ها تست

مشخصات / مجموعه داده	تعداد کل نشست‌ها	تعداد بازدید کننده ناشناخته	تعداد کاربر انسان	تعداد روبات وب غیر مخرب	تعداد روبات وب مخرب
PH	۴۸۴	۸۷	۳۲۱	۲۵	۵۱
PW	۲۷۴۷	۳۲۰	۲۰۷۱	۳۱۵	۴۱

۳- نتایج و شبیه سازی‌ها

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. یک روش ساده برای تخمین خطا بین دسته ها، استفاده از ماتریس تداخل است. ماتریس تداخل، نحوه‌ی توزیع خطا روی دسته های مختلف را مشخص می‌کند. که در جدول ۴ نمونه ای از این ماتریس مشاهده می‌شود. با توجه به این ماتریس نحوه محاسبه صحت و فراخوانی برای کلاس A به ترتیب در فرمول‌های ۱ و ۲ آورده شده است (۲۷).

جدول ۴ : ماتریس تداخل

		کلاس واقعی			
نتیجه پیش بینی شده		A	B	C	D
	A	TP_A	E_{AB}	E_{AC}	E_{AD}
	B	E_{BA}	TP_B	E_{BC}	E_{BD}
	C	E_{CA}	E_{CB}	TP_C	E_{CD}
	D	E_{DA}	E_{DB}	E_{DC}	TP_D

$$\text{Precision}_A = \frac{TP_A}{TP_A + E_{AB} + E_{AC} + E_{AD}} \quad (۱)$$

$$\text{Recall}_A = \frac{TP_A}{TP_A + E_{BA} + E_{CA} + E_{DA}} \quad (۲)$$

$$F_1 = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (۳)$$

برای مجموعه داده پارس هاستینگ با استفاده از روش شبکه باور بیزی و چهارده ویژگی، دقت کلی، میانگین فراخوانی، میانگین صحت و F_1 به ترتیب برابر ۰/۹۵۲۳، ۰/۹۰۵۲، ۰/۸۸۴۳ و ۰/۸۹۴۷ بدست آمده است. ماتریس تداخل برای مجموعه داده پارس هاستینگ در جدول ۵ و دقت دسته بندی در جدول ۶ قابل مشاهده است.

برای مجموعه داده پارس وب سایت با استفاده از روش شبکه باور بیزی و چهارده ویژگی، دقت کلی، میانگین فراخوانی، میانگین صحت و F_1 به ترتیب برابر ۰/۹۵۸۸، ۰/۷۸۷۰، ۰/۷۴۱۲ و ۰/۷۶۳۴ بدست آمده است. ماتریس تداخل برای مجموعه داده پارس وب سایت در جدول ۷ و دقت دسته بندی در جدول ۸ قابل مشاهده است.

جدول ۵ : ماتریس شبکه بیزی برای مجموعه داده پارس هاستینگ

Precision	کلاس واقعی					نتیجه پیش بینی شده
	مخرب وب روبات	وب غیر مخرب روبات	انسان کاربر	ناشناخ کننده		
۰/۹۰۸۰	۰	۰	۸	۷۹	بازدید کننده ناشناخته	
۰/۹۸۷۵	۰	۰	۳۱۷	۴	کاربر انسان	
۰/۷۲	۶	۱۸	۰	۱	روبات وب غیر مخرب	
۰/۹۲۱۶	۴۷	۴	۰	۰	روبات وب مخرب	
	۰/۸۸۶۸	۰/۸۱۸۱	۰/۹۷۵۴	۰/۹۴۰۵	Recall	

جدول ۶ : دقت دسته بندی بر روی مجموعه داده پارس هاستینگ

جمع	روبات وب مخرب	روبات وب غیر مخرب	کاربر انسان	بازدید کننده ناشناخته	
۴۶۱	۴۷	۱۸	۳۱۷	۷۹	مقادیر صحیح دسته بندی شده
۴۸۴	۵۱	۲۵	۳۲۱	۸۷	تعداد نشست‌های واقعی
۰/۹۵۲۴	۰/۹۲۱۵	۰/۷۲	۰/۹۸۷۵	۰/۹۰۸۰	دقت

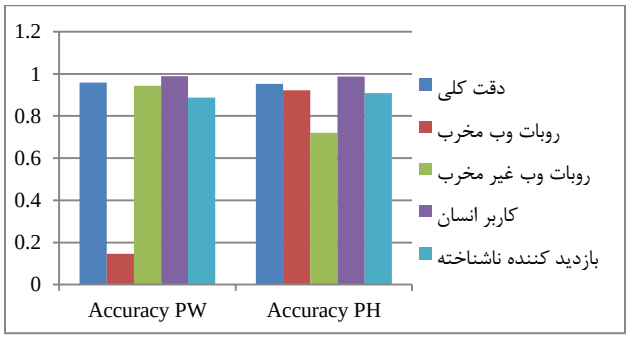
جدول ۷ : ماتریس شبکه بیزی برای مجموعه داده پارس وب سایت

Precision	کلاس واقعی					نتیجه پیش بینی شده
	روبات وب مخرب	روبات وب غیر مخرب	کاربر انسان	بازدید کننده ناشناخته		
۰/۸۸۷۵	۲	۵	۲۹	۲۸۴	بازدید کننده ناشناخته	
۰/۹۸۸۴	۴	۵	۲۰۴۷	۱۵	کاربر انسان	
۰/۹۴۲۸	۸	۲۹۷	۲	۸	روبات وب غیر مخرب	
۰/۱۴۶۳	۶	۴	۳۰	۱	روبات وب مخرب	
	۰/۳	۰/۹۵۵۰	۰/۹۷۱۱	۰/۹۲۲۱	Recall	

جدول ۸ : دقت دسته بندی بر روی مجموعه داده پارس وب سایت

	بازدید کننده ناشناخته	کاربر انسان	روبات وب غیر مخرب	روبات وب مخرب	جمع
مقادیر صحیح دسته بندی شده	۲۸۴	۲۰۴۷	۲۹۷	۶	۲۶۳۴
تعداد نشست‌های واقعی	۳۲۰	۲۰۷۱	۳۱۵	۴۱	۲۷۴۷
دقت	۰/۸۸۷۵	۰/۹۸۸۴	۰/۹۴۲۸	۰/۱۴۶۳	۰/۹۵۸۸

نمودار میله ای مقایسه دقت دسته بندی دو مجموعه داده پارس هاستینگ و پارس وب سایت با روش شبکه باور بیزی در شکل ۳ قابل مشاهده است تعداد نشست‌های استخراج شده از مجموعه داده پارس وب سایت بیشتر از مجموعه داده پارس هاستینگ است و قابل مشاهده است که دقت کلی دسته بندی بر روی مجموعه داده پارس وب سایت بیشتر از پارس هاستینگ می‌باشد.

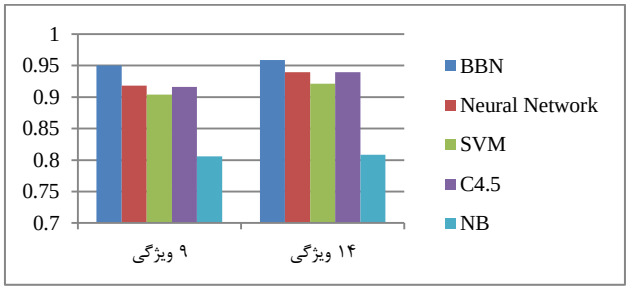


شکل ۳ : نمودار میله ای مقایسه دقت دسته بندی دو مجموعه داده

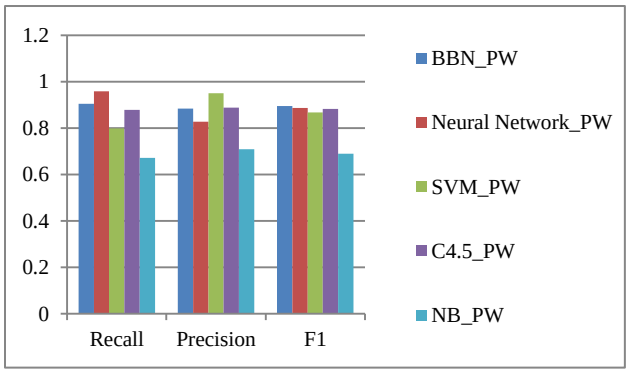
برای ارزیابی دقیق‌تر کارایی با استفاده از چهارده ویژگی بر روی روش‌های شبکه بیزین، شبکه عصبی، SVM و C4.5 که با ویژگی یک تا نه در بخش ۲-۳، انجام شده بودند (۳)، مقایسه صورت گرفت. نتایج این مقایسه در جدول ۹، شکل ۴، ۵ و ۶ قابل مشاهده است. نتایج بدست آمده از ارزیابی‌های مختلف نشان داد که دقت کلی دسته بندی و F_1 با چهارده ویژگی بهتر از سایر روش‌های انجام شده، با نه ویژگی است و همچنین دقت کلی دسته بندی با استفاده از روش شبکه باور بیزی بهتر از سایر روش‌های انجام شده، است.

جدول ۹ : مقایسه دقت کلی دسته بندی با روش‌های مختلف دسته بندی بر روی مجموعه داده پارس وب سایت

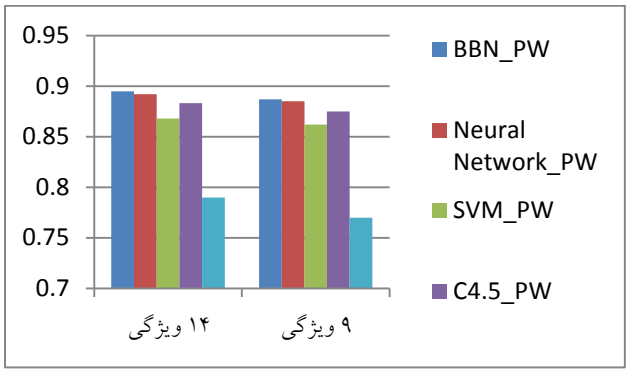
الگوریتم / تعداد ویژگی‌ها	BBN	Neural Network	SVM	C4.5	NB
۹	۰/۹۵	۰/۹۱۸	۰/۹۰۴	۰/۹۱۶	۰/۷۷۶
۱۴	۰/۹۵۸	۰/۹۳۹	۰/۹۲۱	۰/۹۳۹	۰/۸۰۸



شکل ۴ : نمودار میله ای مقایسه دقت کلی دسته بندی با روش‌های مختلف دسته بندی بر روی مجموعه داده پارس وب سایت



شکل ۵: نمودار میله ای مقایسه فراخوانی، صحت و F_1با چهارده ویژگی بر روی مجموعه داده پارس وب سایت با روش‌های دسته بندی



شکل ۶ : نمودار میله ای مقایسه F_1 با چهارده و نه ویژگی

۴- نتیجه گیری

در این مقاله سه ویژگی جدید برای تشخیص بازدیدکنندگان وب سایت‌ها با استفاده از فایل ثبت وقایع مربوط به سرورهای پارس هاستینگ و پارس وب سایت ارائه شده است. دقت دسته بندی برای دو مجموعه داده آموزشی بررسی و مشخص شد که بر روی مجموعه داده پارس وب سایت که تعداد داده های آموزش بیشتر داشت، بالاتر بود.

تشخیص و دسته بندی روبات‌های وبی که تلاش در تقلید رفتار انسان دارند به عنوان مهم‌ترین چالش دسته بندی است، با به‌کارگیری سه ویژگی جدید تا حدودی این چالش کاهش داده شده است. در این مقاله با استفاده از پارامترهای فراخوانی و دقت، روش‌های شبکه بیزین، شبکه عصبی، ماشین بردار پشتیبان، C4.5 و شبکه باور بیزی مورد ارزیابی قرار گرفته است. نتایج بدست آمده از ارزیابی‌های مختلف نشان داد که دقت کلی دسته بندی و F_1 با چهارده ویژگی بهتر از سایر روش‌های انجام شده، با نه ویژگی است و همچنین دقت کلی دسته بندی با استفاده از روش شبکه باور بیزی بهتر از سایر روش‌های انجام شده، است.

۵- 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.
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- Electronics Laboratory, http://www.ife.ee.ethz.ch/education/ WS1_HS2012_02, (2012 December).