۱۰۶۱

تشخیص روبات‌های وب با استفاده از سیستم استنتاج فازی مبتنی بر درخت تصمیم

جواد رجب‌نیا[۱]، مهدیه ذبیحی[۲]، مجید وفایی جهان[۳]

چکیده

روبات‌ها یا خزنده‌های وب برنامه‌هایی برای استخراج دانش از صفحات وب هستند که کار خود را با تعدادی صفحه آغاز کرده و به صورت بازگشتی تمام اسناد قابل دسترسی از این صفحات را بازدید می‌کنند. روبات‌های وب با رفتار های متفاوت اعم از مخرب و غیر مخرب، در کنار کاربران انسانی، جزء بازدیدکنندگان وب به حساب می‌آیند. در این مقاله از یک سیستم استنتاج فازی مبتنی بر درختان تصمیم برای تشخیص نوع بازدیدگنندگان وب استفاده شده است. بازدیدگنندگان به دو دسته انسان و روبات تقسیم می‌شوند. برای تمایز این دو دسته از ۱۴ ویژگی استفاده شده است. کارایی سیستم پیشنهادی در غالب سه معیار صحت، فراخوانی و F1 با سایر روشهای ارائه شده برای تشخیص روبات های وب، ارزیابی می‌شود. علاوه براین از آزمون تی و منحنی Roc برای سنجش نتایج استفاده شده است. نتایج نشان می‌دهد که سیستم پیشنهادی دقت را نسبت به روشهای یادگیری وزن‌دار شده محلی ، طبقه‌بندی کننده Adaboost و Bagging، C4.5 ، شبکه بیزین وشبکه باور بیزی بهبود می بخشد.

کلمات کلیدی

روبات‌های وب، خزنده‌های وب، سیستم استنتاج فازی، درختان تصمیم، فایل ثبت وقایع، کاربرد کاوی وب

[۱] دانشجوی کارشناسی ارشد نرم افزار،دانشگاه امام رضا، مشهد-خیابان دانشگاه- خیابان اسرار-دانشگاه امام رضا،javad.rajabnia@yahoo.com
[۲] دانشجوی کارشناسی ارشد نرم افزار،دانشگاه امام رضا، مشهد-خیابان دانشگاه- خیابان اسرار-دانشگاه امام رضا،mahdieh_zabihi@yahoo.com
[۳] دکتری تخصصی نرم افزار، دانشگاه آزاد اسلامی واحد مشهد، مشهد - قاسم آباد - استاد یوسفی - دانشگاه آزاد اسلامی،vafaeijahan@mshdiau.ac.ir

web robot detection with fuzzy inference system based on decision trees

javad rajabnia; mahdieh zabihi; majid vafaei jahan

ABSTRACT

Web robots or crawlers are programs to extract knowledge from Web pages that traverse the Web autonomously, starting from a ‘‘seed” list of Web pages and recursively visiting documents accessible from that list. Web robots with different behaviors, including destructive and non-destructive, with human users, are considered as website visitors. In this paper, a fuzzy inference system based on decision trees are used to identify website visitors. Visitors are divided into two categories, humans and robots. To distinguish these two sets, 14 features were used. proposed system performance are compared with other methods to detect Web robots by Three criteria in terms of precision, recall and F1. Moreover, the T-test and Roc curve are used to assess the results. The results shows that the proposed system improves accuracy than locally weighted learning, Adaboost, Bagging, C4.5, Bayesian belief networks and Bayesian networks.

KEYWORDS

Web robots, Web crawlers, fuzzy inference system, Decision tree, Access log file, Web usage mining

۱. مقدمه

دنیای امروز دنیایی وابسته به اینترنت است. اینترنت در زمینه‌های مختلف نظامی، آموزش، تجارت، صنعت و بسیاری حوزه‌های دیگر نقش پرکاربردی دارد. مسلما در این بستر وسیع، حجم گسترده‌ای از داده ذخیره می‌شود. نیاز دنیای کنونی، کسب دانش و اطلاعات از این داده‌هاست. برای این منظور از ابزار مختلفی در سطح وب استفاده می‌شود. یکی از این ابزارها، روبات‌های وب نام دارند. روبات‌های وب برنامه‌هایی برای استخراج دانش از صفحات وب هستند که کار خود را از تعدادی صفحه آغاز کرده و به صورت بازگشتی تمام اسناد قابل‌دسترس از این صفحات را بازدید می‌کنند. روبات‌های وب را گاه خزنده[۴]وب، عنکبوت[۵] و یا برداشت‌کننده[۶]می‌نامند. خزنده‌های موتورهای جستجو بر اساس نظمی مشخص به برداشت محتوا از وب می‌پردازند. روبات‌های خریدار[۷] اغلب سایت‌های تجاری را جهت وارسی قیمت‌ها و اجناس خریداری شده پیمایش می‌کنند. خزنده های تمرکزکننده [۸]روبات‌هایی هستند که سراسر وب را جهت یافتن و بررسی صفحات متعلق به یک حوزه معنایی خاص جستجو می‌کنند. خزنده‌های بازبینی[۹] نیز با هدف یافتن لینک‌های خراب یا شکسته به وارسی صفحات وب می‌پردازند. برداشت‌کننده‌ها نوع دیگری از روبات‌های وب هستند که منابعی مانند صفحات اچ تی ام ال، عکس و یا اسناد را از صفحات وب تقاضا می‌کنند(۱۱و۱۷).

این‌ها تنها نمونه محدودی از انواع روبات‌های وب هستند. روبات‌های وب در بستر وب رفتارهای متفاوتی را نشان می‌دهند. برخی از روبات‌های وب رفتاری با قاعده دارند و در مواجه با وب‌سرور براساس پروتکل "robot.txt" عمل کرده و از فرستادن تعداد زیادی درخواست به سرور جلوگیری می‌کنند. در حالیکه برخی از روبات‌های وب دقیقا به صورت هدفمند در طول زمانی کوتاه، حجم وسیعی از تقاضا به یک وب‌سرور می‌فرستند. این روبات‌ها معمولا روبات‌هایی برای حملات توزیعی مختل‌کننده[۱۰] هستند.

این حملات با هدف از کار انداختن سرویس‌های وب سرور، سیل عظیمی از تقاضای توزیعی را به آن ارسال می‌کنند. نوع خاصی از این حملات، معروف به حملات لایه ۷ یا لایه کاربرد، از روبات وب برای پیشبرد اهداف خود استفاده می‌کنند. این می‌تواند دلیل محکمی برای نیاز به تشخیص روبات‌های وب باشد. از دیگر دلایل تشخیص روبات‌های وب میتوان به موارد زیر اشاره کرد. ۱) اغلب روبات‌ها برای استخراج دانش از سایت‌های تجاری اثرات مخربی را در آن به جای می‌گذارند. ۲) روبات‌های وب پهنای باند زیادی را به خود اختصاص داده و سرعت پاسخ دهی سیستم را کاهش می‌دهند. ۳) روبات های وب سهوا یا عمدا اطلاعات وبسایت‌ها را در پایگاه داده مربوط به موتورهای جستجو منتشر می‌کنند(۳).

در این مقاله از یک سیستم استنتاج فازی مبتنی بر درختان تصمیم برای تشخیص و دسته‌بندی بازدیدکنندگان وب استفاده می‌شود. بازدیدکنندگان وب به دو دسته کاربران انسان و روبات وب تقسیم می‌گردند. فایل ثبت وقایع[۱۱] مربوط به سرور پارس وب(۲۷) می‌باشد. از فایل ثبت وقایع ۱۴ خصیصه شناخته شده استخراج می‌شود. هدف، مقایسه دقت سیستم پیشنهادی نسبت به سایر روشهای استفاده شده در این زمینه است. در بخش ۲ کارهای مرتبط معرفی می‌گردند.

۲. کارهای مرتبط

روش‌های متعددی در زمینه تشخیص روبات وب مورد مطالعه قرار گرفته ، که تلاش آنها برای تشخیص و دسته‌بندی بازدیدکنندگان وب با استفاده از فایل ثبت وقایع بوده است. یکی از اولین روشهای ارائه شده توسط کومار و تان در سال ۲۰۰۲ ارائه گردید. آنها دسته بندی بازدیدکنندگان وب را در دو دسته کاربران انسانی و روبات‌های وب با استفاده از درخت تصمیم C4.5 انجام دادند(۹و۱۳). بعد از آن در سال ۲۰۰۳ بلام وهمکارانش با استفاده از تست تورینگ(۱۲)، در سال ۲۰۰۵ بومهارت با استفاده از شبکه بیزین(۲۶)، در سال ۲۰۰۶ وی‌جو و شان با استفاده از زنجیره‌های مارکوف(۲۴)، ولین وهمکارانش با استفاده از روش ترافیک پیوسته در سال ۲۰۰۸(۲۵)، استاسوپولو و دیکایکوس با استفاده از روش شبکه عصبی در سال ۲۰۰۹ به تشخیص روبات‌های وب پرداخته‌اند(۱۱). استوانوویچ با استفاده از روش شبکه عصبی در سال ۲۰۱۳ به دسته‌بندی و تفکیک بازدیدکنندگان وب در چهار دسته روبات‌های مخرب، روبات‌های غیرمخرب، بازدیدکنندگان ناشناخته، کاربران انسانی پرداخته‌اند(۱۰).

---

4 crawler
5 spider
6 harvester
7 Shopping bot
8 Focused crawler
9 verifier
10 DDos
11 Access log

۳. روش شناسایی

مسئله تشخیص روبات وب شامل مراحل ۱) پیش پردازش فایل ثبت وقایع. ۲) تشخیص و استخراج ویژگی‌ها. ۳)برچسب‌گذاری هر نشست (برچسب‌گذاری براساس دو دسته روبات و انسان). ۴) تعیین مدل کلاس‌بندی بازدیدکنندگان وب. ۵) ارزیابی مدل پیشنهادی(۱و۳و۱۱).

۱،۳. پیش‌پردازش

در مسئله تشخیص روبات‌ها، گام پیش‌پردازش شامل تعیین نشست‌های موجود در فایل ثبت وقایع است، که با تقسیم و دسته‌بندی نشست‌ها براساس آدرس آی-پی و اطلاعات عامل کاربر صورت می‌گیرد. هر نشست جدید براساس یک بازه‌ی زمانی مشخص (که در این مقاله ۳۰ دقیقه در نظر گرفته شده‌است) ایجاد می-شود(۳و۴و۵و۶و۱۱).

۱،۱،۳. استخراج ویژگی

در این مقاله، عملیات دسته بندی بازدیدگان وب توسط مدل پیشنهادی با در نظر گرفتن ۱۴ خصیصه بررسی می‌شود که این ویژگی ها به شرح زیر هستند(۱و۳و۱۱):

۱) مدت نشست: به مدت زمان بین اولین و آخرین درخواست گفته می‌شود. هرچه مدت زمان بیشتر باشد بازدیدکننده به روبات نزدیکتر است.

۲) حداکثر نرخ کلیک: به حداکثر تعداد درخواست فایل HTML گفته می‌شود. هر چه میزان این در خواست بیشتر باشد، بازدیدکننده به روبات نزدیک‌تر است.

۳) درخواست فایل "robot.txt" : روبات‌های شناخته شده معمولا برای درخواست اطلاعات و دسترسی به یک سایت ابتدا این فایل را درخواست می‌کنند، به عنوان نمونه یک روبات برای درخواست اطلاعات از سایت

"www.google.com" ابتدا "www.google.com/robot.txt" را درخواست می‌دهد.(برچسب‌گذاری نشست)

۴) درصد درخواست HTTP از نوع Head: روبات‌های وب برای کاهش حجم اطلاعات مربوط به درخواست، از ارسال یک عنوان به سرور استفاده می‌کنند. این درحالیست که کاربران انسانی با استفاده از مرورگرها درخواست خود را به سرور ارسال می‌کنند، که در این روش متد درخواست GET است.

۵) درصد درخواست با ارجاعات خالی: یک خصیصه عددی است که درصد درخواست‌ها با ارجاع خالی در یک نشست کاربر را نشان می‌دهد. این خصیه برای اکثر روبات ها مقدار بالایی دارد چراکه مرورگرهای وب اغلب اطلاعاتی را به عنوان ارجاعات به صورت پیش‌فرض مقداردهی می‌کنند.

۶) میزان 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 ندارند، 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.

۷) حجم اطلاعات درخواستی: این خصیصه عددی حجم اطلاعات درخواستی در یک نشست را برحسب بایت بیان می‌کند که هرچه بیشتر باشد بازدیدکننده به روبات وب نزدیک‌تر است.

۸) نرخ درخواست دنباله متوالی: یک خصیصه عددی است که درصد درخواست‌های متوالی برای صفحات متعلق به یک مسیر یکسان از وب را در طول یک نشست نشان می‌دهد. برای مثال درخواست "/google/translate/*.*" به‌عنوان یک دنباله متوالی از درخواست‌های HTTP در نظر گرفته می‌شود.

۹) پاسخ خطای ۴۰۴: روبات های وب نسبت به کاربران انسانی دارای درصد بالاتری از انتخاب لینک‌های خراب هستند. (برای برچسب‌گذاری نشست‌ها) .

۱۰) عمق درخواست صفحه: یک خصیصه عددی است که عمق صفحه در تمام درخواست‌های موجود در یک نشست را نشان می‌دهد. برای مثال درخواست "/google/translate/persianpage.html" با عمق ۳ تعیین می‌شود. و درخواست "/google/translate.html" با عمق ۲ در نظر گرفته می‌شود.

۱۱) درصد درخواست فایل PDF و PSS: روبات‌های وب درخواست‌های بیشتری برای فایل های PDF و Postscript جهت جمع‌آوری اطلاعات دارند.

۱۲) نسبت درخواست Html به تصویر: روبات‌های وب نسبت به کاربران انسانی فایل Html بیشتری را درخواست می‌دهند، برعکس این، کاربران انسانی درخواست بیشتری برای تصویر دارند، بنابراین هر چه این نسبت بیشتر باشد بازدیدکننده به روبات نزدیک‌تر است.

۱۳) درصد 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 است. بنابراین اگر صفحه وبی بدون مشاهده تمام منابع جاسازی شده در آن، در یک نشست مشاهده شود، آن نشست مشکوک به روبات وب است.

۱۴) درصدی از کوکی‌ها : کوکی‌ها اطلاعاتی هستند که سرور HTTP می‌تواند به همراه منبع درخواست شده به ماشین کاربر ارسال کند. کاربر ممکن است این اطلاعات را ذخیره کند و متعاقبا هنگام ارسال درخواست‌های بعدی اطلاعات آن را به سرور پس بفرستد. اگر درصد کوکی‌ها در یک نشست صفر باشد نشست مشکوک به روبات است.

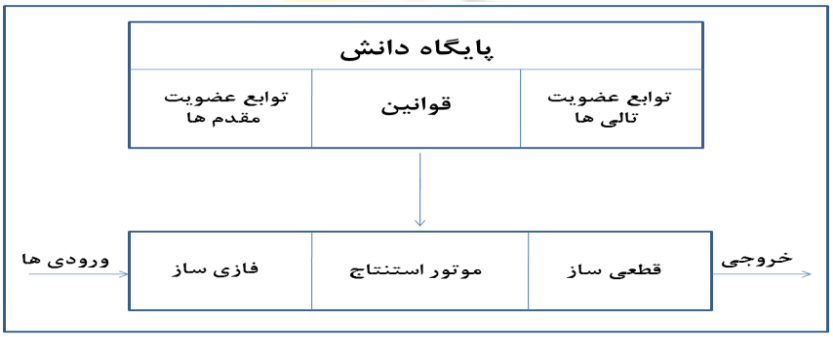
۲،۱،۳. برچسب‌گذاری نشست‌ها

پس از تعیین خصیصه‌ها برای هر نشست، نوبت به برچسب‌گذاری نشست‌ها می‌رسد. در مرحله اول تمام نشست‌هایی که خصیصه"robot.txt" آنها مقدار یک است، به عنوان روبات در نظر گرفته می‌شوند. در ادامه رشته عامل کاربر کلیه نشست‌های باقیمانده با لیست رشته عامل کاربر روبات‌های شناخته شده وب مقایسه می-شود. در صورتیکه رشته عامل کاربر با لیست بروزشده عامل‌های کاربر روبات‌های وب مطابقت داشت، آن نشست نیز به عنوان روبات وب شناخته می‌شود. اگر رشته عامل کاربر با لیست بروزشده مرورگرهای وب مطابقت داشته باشد و فایل "robot.txt" در این نشست خوانده نشده باشد، نشست به عنوان انسان برچسب‌گذاری می-شود.

۴. مدل پیشنهادی

۱،۲.سیستم استنتاج فازی

استنتاج فازی فرایند فرموله‌کردن یک نگاشت از ورودی‌های داده شده به یک خروجی، با استفاده از منطق فازی است. این نگاشت، اساسی برای گرفتن یک تصمیم یا تشخیص یک سری الگو خواهد بود(۲۰). ساختار این سیستم در شکل زیر آمده‌است(۲۱و۲۲):



شکل ۱. ساختار کلی یک سیستم استنتاج فازی

توصیفی از اجزای شکل (۱):

پایگاه دانش: هسته سیستم که شامل کلیه قوانین سیستم و تعریف توابع عضویت مربوط به مقدم و تالی هر یک از قوانین است.

موتور استنتاج: که بر اساس یک مکانیزم از پیش‌تعریف شده، اطلاعات پایگاه دانش را به کار می‌برد.

فازی‌ساز و قطعی‌ساز: دو بلاک که به ترتیب موتور استنتاج را به ورودی‌ها و خروجی سیستم متصل می‌کنند.

در این مقاله، ورودی‌های سیستم، ۱۴ خصیصه مربوط به تشخیص روبات، و خروجی نیز متغیری است که انسان یا روبات بودن هر نشست را تعیین می‌کند. علاوه بر این نوع سیستم استناج فازی بکار رفته ممدانی است. در این سیستم، استنتاج از پنج مرحله اصلی زیر تشکیل می‌شود:

۱- فازی‌سازی ورودی‌ها

۲- به کار بردن عملگرهای فازی. در این مقاله متد مورداستفاده برای عملگر AND، متد حداقل‌ساز[۱۲] و برای عملگر OR، متد حداکثرساز[۱۳] است.

۳- استفاده از متد استنباط[۱۴]. در این مقاله از متد حداقل‌ساز برای استنباط استفاده شده است.

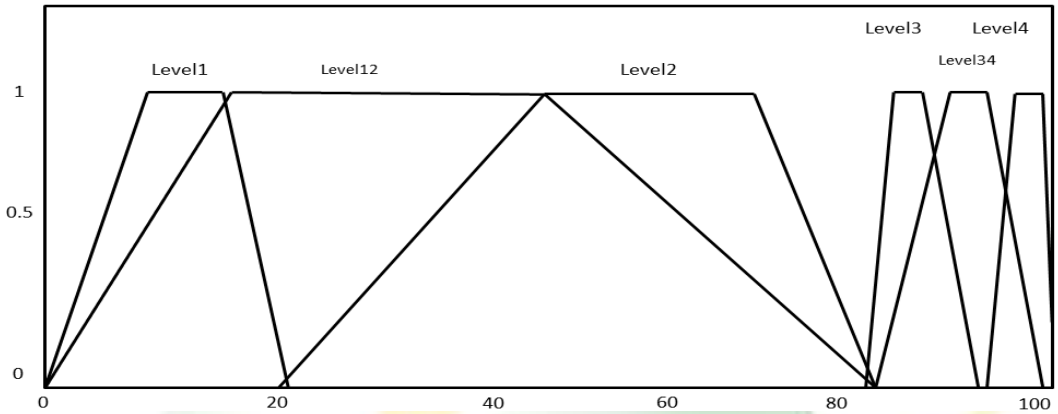
۴- یکی‌کردن تمام خروجی‌ها[۱۵]. در این مقاله، از تابع حداکثرساز برای این کار استفاده شده است.

[12] Min method
[13] Max method
[14] Implication method
[15] Aggregation method

۲. قطعی‌سازی[۱۶]. در این مقاله از متد "centroid" استفاده شده است.

در روش پیشنهادی، جهت آموزش و ایجاد توابع عضویت برای هر یک از ۱۴ خصیصه، از درخت تصمیم C4.5 استفاده می‌شود. ابتدا با استفاده از مجموعه داده موجود، درخت تصمیم ساخته می‌شود. سپس از درخت حاصل برای استخراج قوانین و نقاط شکست استفاده می‌شود. به طوریکه به ازای هر خصیصه نقاط شکست متفاوتی بدست می‌آید. در ادامه، با داشتن نقاط شکست، توابع عضویت مربوط به هر خصیصه مشخص می‌گردد. شکل(۲) نمونه ای از این توابع را نشان می‌دهد.

پس از تعریف و رسم توابع عضویت، قواعد استخراج شده از درخت بعنوان قوانین سیستم استنتاج فازی ثبت می‌شوند و بدین ترتیب مرحله آموزش بر روی سیستم پیشنهادی اعمال می‌گردد. در مرحله آزمون، مجموعه داده تست بعنوان ورودی به سیستم داده می‌شود. خروجی نیز عددی بین ۰ تا ۱ است. در این مقاله نقطه شکست برای تابع عضویت خروجی برابر ۰,۴ در نظر گرفته شده‌است که بیشترین دقت را در تشخیص نوع بازدیدکنندگان وب دارد.



شکل ۲. نمودار مربوط تابع عضویت خصیصه "نرخ درخواست متوالی"

[16] Defuzzification

مشخصات / مجموعه داده	تعدادکل نشست ها	تعداد نشست از نوع انسان	تعداد نشست از نوع روبات	میانگین دقت	میانگین صحت	میانگین نرخ فراخوانی	میانگین F1
PW_Ds1	۲۸۴۶	۱۱۸۱	۱۱۸۱	۹۸,۹۷	۹۹,۲۴	۹۹,۰۷	۹۹,۱۵
PW_Ds2	۲۴۲۲	۷۵۷	۱۱۸۱	۹۸,۷	۹۹	۹۹,۰۵	۹۹,۰۲
PW_Ds3	۱۹۹۸	۷۵۷	۷۵۷	۹۷,۸۴	۹۷,۹۵	۹۸,۰۲	۹۷,۹۸

جدول ۱. مشخصات مجموعه داده‌ها

۵. مقایسات و نتایج

آزمایشات انجام‌شده در این مقاله بر روی فایل ثبت وقایع سرور پارس‌وب انجام گرفته‌است. مجموعه داده حاصل به نام PW_Ds2 دارای ۲۴۲۲ نشست شامل ۷۵۷ انسان و ۱۱۸۱ روبات می‌باشد. برای رفع مشکل عدم توازن کلاس‌ها از تولید مجدد نمونه‌ها۱۷ استفاده شده‌است(۱۱). در PW_Ds1 نرخ انسان به ۵۰ درصد رسیده و با نرخ روبات برابر شده‌است. و در PW_Ds3 نرخ نمونه روبات کاهش یافته و به نرخ انسان رسیده است. در ادامه روش پیشنهادی با شش روش داده کاوی مقایسه می‌شود.

۱,۵. پارامترهای ارزیابی سیستم

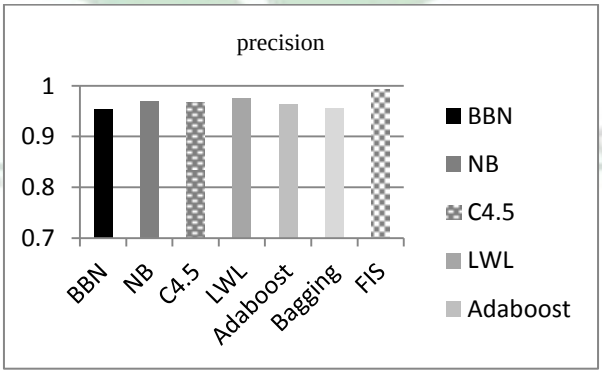
در این بخش به معرفی پارامترهای ارزیابی مورد نیاز برای سیستم به صورت خلاصه پرداخته می‌شود(۳و۱۱). برای تست صحت و کارایی، از متریک های بازیابی اطلاعات مانند *Recall*، *Precision* و *F-Measure* استفاده شده‌است. با توجه به جدول (۱) مقدار این پارامترها برای PW_Ds1 از سایرین بیشتر است، پس در ادامه کلیه آزمایشات بر روی همین مجموعه‌داده انجام گرفته می‌شود.

۱) Precision: یک متریک عمومی که برای اندازه‌گیری مفید بون[۱۸] الگوریتم پیشنهادی به کار می‌رود و به صورت معادله زیر می‌باشد.

$$precision = \frac{tp}{tp+fp} \quad (۱)$$

در تعیین صحت پیش‌بینی روبات، رابطه بالا به این صورت تعریف می‌شود که tp تعداد کل نشست‌های از نوع روبات است که بدرستی پیش‌بینی شده‌اند و fp تعداد نشست‌های انسانی است که اشتباها روبات شناسایی شده‌اند.

نمودار زیر میانگین صحت روش پیشنهادی در مقایسه با سایر روشها را نشان می‌دهد.



شکل ۳. نمودار صحت روشهای مختلف در تشخیص روبات وب

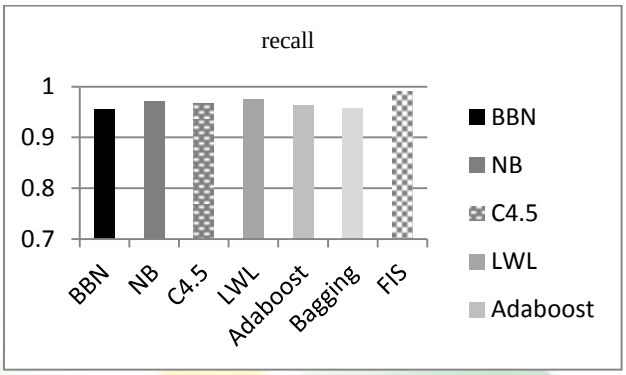
[17] Resampling
[18] Usefulness

۲) Recall: یک متریک عمومی است که برای ارزیابی سودمند بودن[19] الگوریتم پیشنهادی به کار می‌رود و به صورت معادله زیر عمل می‌نماید.

$$recall = \frac{tp}{tp+fn} \quad (۲)$$

در این رابطه نیز برای نشست‌های روبات، fn تعداد کل نشست‌های روباتی است که اشتباها انسان پیش‌بینی شده‌اند.

نمودار زیر میانگین نرخ فراخوانی روش پیشنهادی در مقایسه با سایر روشها را نشان می‌دهد.



شکل ۴.نمودار نرخ فراخوانی روشهای مختلف در تشخیص روبات وب

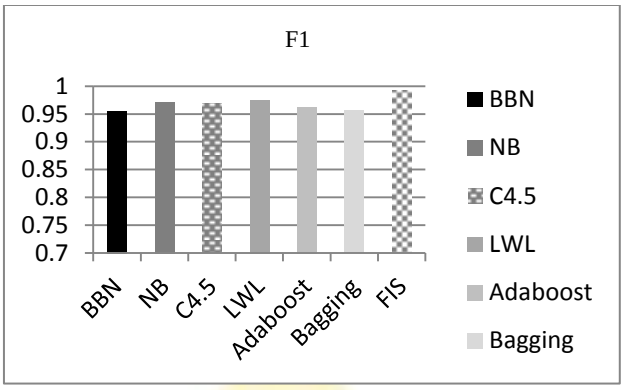
۳) F1: یکی دیگر از متریک‌های ارزیابی است که با استفاده از پارامترهای *Recall* و *Precision* به صورت رابطه‌ی (۳) به دست می‌آید.

رابطه زیر، رابطه *F1* است که ترکیبی از دو معیار بالاست. برای اینکه *F1* مقدار بزرگی داشته باشد، باید صحت و نرخ فراخوانی مقادیر نزدیک به هم داشته باشند. وگرنه *F1* به سمت مقدار کوچکتر متمایل خواهد شد.

$$F1 = \frac{2\times precision\times recall}{precision+recall} \quad (۳)$$

نمودار زیر معیار فوق را برای هر یک از روش‌ها نشان می دهد.

[19] Utility



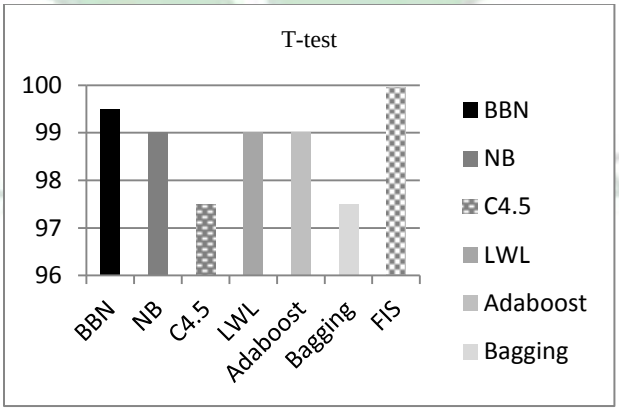
شکل ۵. نمودار F1 روشهای مختلف در تشخیص روبات وب

۲،۵. آزمون تی

راه دیگر برای ارزیابی یک روش کلاس‌بندی، استفاده از آزمون تی است(۳و۲۳). بعد از اعمال روش کلاس‌بندی مربوطه نشست‌ها به دوگروه *tp* [۲۰]و *tn*[۲۱] تقسیم می‌شوند. نشست‌هایی در گروه *tn* دسته بندی می‌شوند که انسان بوده و به درستی پیش‌بینی شده‌اند. نشست‌های گروه *tp* نیز روبات هایی هستند که درست شناسایی شده‌اند. در ادامه میانگین و انحراف معیار خصیصه نرخ درخواست دنباله متوالی در هر دو گروه محاسبه شده و طبق رابطه (۴) مقدار آزمون تی مشخص می‌گردد.

$$t = \frac{|mean1(f) - mean2(f)|}{\sqrt{\frac{Var1}{n1} + \frac{Var2}{n2}}} \qquad (۴)$$

این کمیت برای سیستم پیشنهادی برابر ۶۵،۵۳ است. در این رابطه $mean_1$ و $mean_2$ میانگین خصیصه نرخ دنباله متوالی در گروه‌های ۱ و۲، و var_1 و var_2 انحراف معیار این خصیصه در گروه‌های مذکور است. n_1 و n_2 تعداد نشست‌های موجود در هر دوگروه را نشان می‌دهند. درجه آزادی این تست از رابطه n_1+n_2-1 محاسبه می‌شود. با توجه به اعداد حاصل از آزمون، یعنی کمیت t و درجه آزادی محاسبه شده سیستم استنتاج فازی آزمون تی را با دقت ۹۹،۹۵ پاس می‌کند. نمودار زیر نتایج حاصل از این آزمون بر روی روش‌های مختلف را نشان می‌دهد.



شکل ۶- نمودار مربوط به نتایج تست تی

[20] True positive
[21] True negative

۳,۵. منحنی Roc

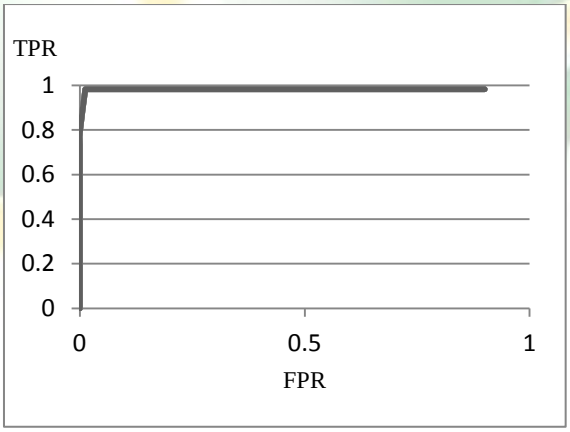
منحنی Roc [۲۲]برای نمایش تعادل مابین کمیت‌های TPR[۲۳] و FPR[۲۴] در یک روش کلاس‌بندی استفاده می‌شود(۱۹و۱۸). TPR کسری از نمونه‌های روبات است که به درستی پیش‌بینی شده اند، در حالی که FPR کسری از نمونه‌های انسانی است که به صورت روبات تشخیص داده‌شده‌اند.

روابط زیر برای محاسبه این دو کمیت استفاده می‌شود.

$$TPR = \frac{True\ positive}{True\ positive + False\ negative} \qquad (۵)$$

$$\mathrm{FPR} = \frac{False\ positive}{\mathrm{True\ negative + False\ positive}} \qquad (۶)$$

در منحنی Roc، FPR محور xها و TPR محور yها را نشان می‌دهد. پس هر نقطه این منحنی، نشان دهنده زوجی به صورت (FPR,TPR) است. برای بدست آوردن این نقاط، یک حدآستانه متمایزکننده برای روش کلاس‌بندی تعیین می‌گردد و هر بار نرخ FPR و TPR برای روش کلاس‌بندی مربوطه با توجه به این حد آستانه، محاسبه می‌شود. درنهایت تمام نقاط حاصل با خطوطی به هم متصل شده و منحنی شکل می‌گیرد. در این مقاله، حدآستانه با تغییر مقادیر تابع عضویت خروجی سیستم فازی می‌گردد. یک منحنی Roc مطلوب، تا حدامکان به نقطه (۰,۱) نزدیک است. جائیکه FPR مقدار صفر و TPR مقدار یک را دارد. به این معنا که مدل ارائه شده برای کلاس‌بندی، هیچ انسانی را به عنوان روبات شناسایی نکرده‌است و تمام روبات‌های موجود در نمونه بدرستی به عنوان روبات تشخیص داده شده‌اند. نقطه (۰,۰) نشان دهنده حد آستانه‌ای است که در آن تمام نمونه‌ها به عنوان انسان پیش-بینی می‌شوند و برعکس نقطه (۱,۱) حد آستانه‌ای را معرفی می‌کند که باعث پیش‌بینی تمام نمونه‌ها به عنوان روبات می‌گردد. هر چه نمودار با شیب بیشتری به سمت نقطه (۰,۱) متمایل گردد مطلوب‌تر است. مشخصه مهم این نمودار مساحت زیر شکل حاصل از آن است که به آن AUC[۲۵] گویند. کمیت AUC برای یک کلاس‌بندی خوب برابر یک است.



شکل۷- منحنی ROC سیستم پیشنهادی

کمیت AUC برای سیستم پیشنهادی برابر ۰,۹۸۳۳ است.

[22] Receiver Operating Characteristic
[23] True positive rate
[24] False positive rate
[25] area under the curve

۶. نتیجه‌گیری

در این مقاله یک سیستم استنتاج فازی مبتنی بر درختان تصمیم طراحی شده است که به تشخیص روبات‌های وب می‌پردازد. روش پیشنهادی بر اساس معیارهای ارزیابی صحت، نرخ فراخوانی و کمیت *F1* با روشهای یادگیری وزن‌دار شده محلی، طبقه‌بندی کننده Adaboost و Bagging، C4.5 ، شبکه بیزین و شبکه باور بیزی مقایسه شده‌است. نتایج حاصل نشان می‌دهد که هر سه معیار در سیستم استنتاج فازی بالاتر از روش های نامبرده است و روش پیشنهادی این مقاله دارای دقت بالاتری برای تمایز روبات‌های وب از کاربران انسانی می‌باشد. علاوه بر این با ایجاد توازن در تعداد نمونه‌های روبات و انسان موجود در مجموعه آموزشی می‌توان نرخ سه معیار فوق را افزایش داد. در ادامه از آزمون تی نیز برای مقایسه روش پیشنهادی با سایر روش‌های مزبور استفاده شده‌است. با توجه به نتایج، سیستم استنتاج فازی با دقت بالاتری برابر ۹۹,۹۵ درصد این آزمون را پاس می‌کند. در نهایت رسم منحنی Roc نشان می‌دهد که مساحت زیر این نمودار برابر ۰,۹۸۳۳ است که بر دقت این سیستم در تمایز روبات وب از کاربر انسانی دلالت دارد.

۷. مراجع

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

[2] 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.

[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] S. Layeghi, A.H. Zarei, M. Vafaei Jahan, M. Jalali, "*Neuro-Fuzzy Method for Malicious Web Robot Detection*" Computer Science and Engineering Conference, Islamic Azad University, Najaf Abad Branch, 2013.

[5] S. Layeghi, A.H. Zarei, M. Vafaei Jahan, M. Jalali, "*The Analysis of the Data Mining Methods for Malicious Web Robot Detection,*" National Conference on Application of Intelligent Systems in Science and Technology, Islamic Azad University, Quchan Branch, 2013.

[6] A.H. Zarei, S. Layeghi, M. Vafaei Jahan, M. Jalali, "*Bayesian Believe Network for Malicious Web Robot Detection,*" The 18th International Conference on Computer Association, Sharif University of Technology, Tehran, Iran, 2013.

[7] 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.

[8] 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.

[9] 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.

[۱۰] 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.

[11] 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.

[12] 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.

[13] Tan P.-N., Kumar V. "*Modeling of Web robot navigational patterns*". In Workshop on Web Mining for E-Commerce. Challenges and Opportunities Working Notes (KDD2000), Boston, MA; August 2000. p. 111–17.

[14] Kharwar, A., Kapadia, V., "*A Complete PreProcessing Method For Web Usage Mining*", GANPAT university journal of engineering & technology, Vol.1, issue 1, march 2011.

[15] Known, S., Oh, M., Kim, D., Lee, J., Kim,Y., Cha,S. ,"*Web Robot Detection Based on Monotonous Behavior*", Springer-Verlag Berlin Heidelberg 2012.

[16] Losawar,V.Joshi,M.,"*Data Preprocessing in Web Usage Mining*", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore.

[17] Doran, D., Swapna S. Gokhale , *"A Classification Framework for Web Robots"*, journal of the american society for information science and thechnology, December 2012.

[18] Provost, F., Fawcett, T., " *Analysis and visualization of classifier performance: Comparison under Imprecise class and cost distribution*" , Third International Confer_
ence on Knowledge Discovery and Data Mining (KDD-97)
Huntington Beach, CA, pp.43-48, 1997.

[19] Batista, G., Prati, R., Monard, M., " *A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data* " , Sigkdd Explorations, Vol.6, Issue 1, pp.20-29, 2004.

[20] Sumalatha, V., Ramani, K., Lakshmi, K., " *Fuzzy Inference System to Control PC Power Failures* " , International Journal of Computer Applications (0975 – 8887), Vol. 28– No.4, pp.10-17, August 2011.

[21] Siler, W., Buckley, J., "*Fuzzy Expert Systems and Fuzzy Reasoning*: JOHN WILEY & SONS,INC.".

[22] Ross, T. (2010). "*Fuzzy Logic with engineering applictions*: JOHN WILEY & SONS,INC."

[23] Kanji, G. (2006). "100 Statistical Tests: STAGE".

[24] 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.

[25] Lin, X., Quan, L., Wu, H., "*An Automatic Scheme to Categorize User Sessions in Modern HTTP Traffic*", in Proc. GLOBECOM, pp.1485-1490, 2008.

[26] 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.

[27] Pars Web Site. November 2012, http://www.parswebsite.com.

[28] http://www.user-agents.org/, August 2011.

[29] http://www.botsvsbrowsers.com, August 2011.

[30] http://user-agent-string.info/list-of-ua/bots-ip, 2012.