تشخیص روبات‌های وب با استفاده از نظریه مجموعه‌های فازی ناهموار

سمانه رحیمی و جواد حمیدزاده

چکیده: روبات‌های وب، برنامه‌های نرم‌افزاری هستند که به طور خودکار در اینترنت اجرا می‌شوند و مهم‌ترین وظیفه آنها واکشی اطلاعات و ارسال آنها به سرویس‌دهنده مبدأ است. مصرف زیاد پهنای باند شبکه توسط آنها و کاهش کارایی سرویس‌دهنده باعث شده تا مسأله تشخیص روبات‌های وب مطرح شود. در این مقاله از نظریه مجموعه‌های فازی ناهموار برای تشخیص روبات‌های وب استفاده شده است. روش پیشنهادی شامل چهار مرحله است. در مرحله اول، نشست‌های کاربران وب توسط خوشه‌بندی مجموعه‌های فازی ناهموار شناسایی می‌شود. در مرحله دوم، برداری شامل ۱۰ ویژگی متمایز برای هر نشست استخراج می‌گردد. در مرحله سوم نشست‌های شناسایی‌شده توسط یک روش مکاشفه‌ای برچسب‌گذاری می‌شود. در مرحله چهارم این برچسب‌ها با استفاده از طبقه‌بندی مجموعه‌های فازی ناهموار بهبود می‌یابد. کارایی روش پیشنهادی بر روی مجموعه داده‌های واقعی ارزیابی شده است. نتایج آزمایش‌ها نشان‌دهنده برتری روش پیشنهادی نسبت به سایر روش‌های مطرح از نظر معیار F است.

کلیدواژه: **پیش‌پردازش فایل ثبت وقایع، تشخیص روبات‌های وب، شناسایی نشست‌های بازدیدکنندگان وب، نظریه مجموعه‌های فازی ناهموار.**

۱- مقدمه

امروزه اینترنت یکی از گسترده‌ترین فناوری‌هایی است که در جهان واقعی توسعه یافته است. وجود اطلاعات متنوع و برنامه‌های کاربردی مختلف در بستر آن باعث شده است تا دسته دیگری از بازدیدکنندگان اینترنت به نام روبات‌های وب[۱] در حال پیمایش و مرور آن باشند. روبات‌های وب، برنامه‌های نرم‌افزاری هستند که به طور خودکار، درخواست‌هایی را برای گرفتن منابع سرویس‌دهنده به آن ارسال می‌کنند و به واکشی اطلاعات، تحلیل آنها و ارسال نتایج به سرویس‌دهنده مبدأ می‌پردازند. به روبات‌های وب، خزنده‌های وب[۲]، سرگردان‌های وب[۳]، برداشت‌کننده‌های وب[۴] و عنکبوت‌های وب[۵] نیز گفته می‌شود [۱] و [۲]. خزنده‌های موتور جستجو [۳]، روبات‌های خریدار، خزنده‌های متمرکز (برداشت‌کننده‌ها)، روبات‌های بررسی‌کننده پیوندها و روبات‌های جمع‌آوری آمار سایت، نمونه‌ای از روبات‌های خوش‌رفتار اینترنت هستند [۱]، [۴] و [۵]. روبات‌های حمله‌کننده DDOS و خزنده‌های جمع‌کننده ایمیل‌های کاربران نیز، نوع رایجی از روبات‌های مخرب هستند [۵] و [۶]. مطالعات اخیر نشان داده است که بیش از ۵۰ درصد بازدیدها، مربوط به ترافیک روبات‌های وب است و همچنین میزان قابل توجهی از پهنای باند شبکه را مصرف می‌کنند. از طرف دیگر، وجود روبات‌های مخرب ممکن است امنیت کاربران انسانی را به خطر بیاندازد [۱]. برای جلوگیری از اثرات سوء روبات‌های مخرب و همچنین قابلیت برنامه‌ریزی ظرفیت یک سرویس‌دهنده، لازم است ترافیک روبات‌های وب شناسایی شود.

هر سرویس‌دهنده وب یک فایل ثبت وقایع دارد و درخواست‌های بازدیدکنندگان مختلف به عنوان یک رکورد در این فایل ثبت می‌شود. این فایل در پژوهش‌های حوزه کاوش وب مورد تحلیل قرار می‌گیرد.

تکنیک‌های تشخیص روبات‌های وب به دو دسته درون‌خط[۶] و برون‌خط[۷] تقسیم می‌شوند [۷] و تاکنون پژوهش‌های زیادی در این حوزه انجام شده است [۸] تا [۱۴]. تکنیک‌های برون‌خط، فایل ثبت وقایع سرویس‌دهنده وب را بعد از اتمام پیمایش‌های بازدیدکنندگان تحلیل می‌کنند و اکثر آنها رکوردهای درخواست این فایل را به مجموعه نشست‌هایی تقسیم می‌کنند. هر نشست مجموعه درخواست‌هایی است که از طرف یک بازدیدکننده خاص ارسال شده است. دو ویژگی در رکوردهای درخواست که در مرور ادبیات تشخیص روبات‌های وب حائز اهمیت است، آدرس پروتکل اینترنتی[۸] و نام عامل کاربری درخواست‌دهنده[۹] است. اکثر مقالات این حوزه مانند [۲]، [۴] تا [۶]، [۱۴] و [۱۵] از این ویژگی‌ها برای شناسایی نشست‌ها و برچسب‌گذاری آنها استفاده می‌کنند. وجود نرم‌افزارهای پنهان‌کننده هویت مانند انواع فیلترشکن‌ها و همچنین استفاده از نماینده سرویس‌دهنده[۱۰] مشابه، باعث می‌شود درخواست‌هایی با نام عامل کاربری مشابه با آدرس‌های پروتکل اینترنتی مختلفی در فایل ثبت وقایع ذخیره شوند و یا درخواست‌های دارای یک آدرس پروتکل اینترنتی، نام عامل کاربری مختلفی داشته باشند. بنابراین یک عدم قطعیت ذاتی[۱۱] در روش‌های مبتنی بر آدرس پروتکل اینترنتی[۱۲] وجود دارد [۴]. همچنین از آنجایی که پروتکل http، یک پروتکل بدون اتصال[۱۳] و بدون وضعیت[۱۴] است، این رکوردهای درخواست ممکن است حاوی اطلاعات ناکامل[۱۵] باشند [۱۲]. از طرف دیگر بعضی از روبات‌های وب سعی می‌کنند پیمایش‌های خود را مخفی نگه دارند. آنها از اطلاعات و

این مقاله در تاریخ ۳ اسفند ماه ۱۳۹۴ دریافت و در تاریخ ۲۳ فروردین ماه ۱۳۹۶ بازنگری شد.

سمانه رحیمی، گروه مهندسی کامپیوتر، دانشگاه بین‌المللی امام رضا علیه‌السلام، مشهد، (email: samane.rahimi@imamreza.ac.ir).

جواد حمیدزاده، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی سجاد، مشهد، (email: j_hamidzadeh@sadjad.ac.ir).

1. Web Robots
2. Web Crawlers
3. Web Wanderers
4. Web Harvesters
5. Web Spiders
6. Online
7. Offline
8. IP Address
9. User Agent String
10. Proxy Server
11. Inherent Uncertainty
12. IP-Based Techniques
13. Connection-Less Protocol
14. Stateless Protocol
15. Incomplete Data

الگوهای پیمایشی کاربران انسانی تقلید می‌کنند تا به راحتی شناسایی نشوند. بنابراین با توجه به بحث‌های فوق، چالش‌های مسأله تشخیص روبات‌های وب را می‌توان به صورت زیر در نظر گرفت:

- عدم اطمینان روش‌های مبتنی بر آدرس پروتکل اینترنتی.
- تلاش روبات‌های وب در مخفی نگه داشتن اطلاعات خود و تقلید از رفتارهای کاربران انسانی.
- عدم استفاده از شباهت درخواست‌های بازدیدکنندگان از نقطه نظر همه ویژگی‌های تعریف‌شده.

در این مقاله سعی شده تا با استفاده از نظریه مجموعه‌های فازی ناهموار[1] [۱۶] که برای مقابله با داده‌های مبهم و ناکامل ارائه شده است، چالش‌های فوق تا حدی برطرف شوند. این نظریه از مفاهیم قوی ریاضی استفاده می‌کند و تاکنون در مسایل حوزه کاوش وب [۱۷]، بهبود بردار پشتیبان توصیف داده‌ها [۱۸]، انتخاب ویژگی [۱۹] و [۲۰] و انتخاب نمونه [۲۱] و [۲۲] استفاده شده است. یکی از پژوهش‌های اخیر نیز از مجموعه‌های فازی ناهموار برای انتخاب ویژگی روبات‌های وب استفاده کرده است [۲۳]. نویسندگان مشابه اکثر پژوهش‌های مرتبط، نشست‌های بازدیدکنندگان را توسط دو ویژگی آدرس پروتکل اینترنتی و نام عامل کاربری درخواست‌ها شناسایی کرده‌اند. بنا بر چالش‌های مطرح‌شده، نشست‌های شناسایی‌شده از اطمینان قابل قبولی برخوردار نیستند.

در این مقاله در حوزه تشخیص روبات‌های وب برای اولین بار شناسایی نشست‌های بازدیدکنندگان توسط خوشه‌بندی مجموعه‌های فازی ناهموار بر اساس فاصله جارو- وینکلر[2] انجام شده است. در روش پیشنهادی، ابتدا رابطه بین رکوردهای درخواست با استفاده از معیار شباهت رشته‌ای تعیین شده و سپس از این رابطه برای گروه‌بندی درخواست‌ها (شناسایی نشست‌ها) استفاده شده است[3]. در ادامه برای هر نشست برداری شامل ۱۰ ویژگی عددی استخراج شده است[4] و در انتها نشست‌ها توسط یک روش مکاشفه‌ای[5] برچسب‌گذاری شده‌اند[6]. همچنین برای اطمینان از صحت برچسب‌های تخصیص داده شده، این برچسب‌ها توسط طبقه‌بندی مجموعه‌های فازی ناهموار بهبود یافته‌اند.

بنابراین هدف این مقاله تشخیص دقیق‌تر روبات‌های وب است و نوآوری مقاله در دو قسمت است: ۱) شناسایی نشست‌های بازدیدکنندگان توسط خوشه‌بندی مجموعه‌های فازی ناهموار و ۲) بهبود برچسب نشست‌ها توسط طبقه‌بندی مجموعه‌های فازی ناهموار.

بعضی از پژوهش‌های حوزه کاوش وب مانند [۲۴] و [۲۵] نیز برای شناسایی نشست‌ها از مجموعه‌های فازی ناهموار استفاده کرده‌اند. تفاوت این مقاله و پژوهش‌های ذکرشده این است که اولاً آنها در حوزه مسایل کاوش وب ارائه شده‌اند. این مسایل برخلاف مسأله تشخیص روبات‌های وب، ابتدا درخواست‌های مشکوک به روبات وب را حذف می‌کنند. در ادامه کاربران متفاوت و نشست‌های هر کدام را شناسایی کرده و الگوهای مناسب برای کاربران را استخراج می‌کنند. ثانیاً مقالات ذکرشده ابتدا نشست‌ها را طبق مقایسه آدرس پروتکل اینترنتی و نام عامل کاربری شناسایی می‌کنند و سپس در مرحله بعد، شباهت این نشست‌های شناسایی‌شده را با هم به دست آورده و آنها را خوشه‌بندی می‌کنند. به عبارت دیگر نشست‌های شناسایی‌شده را توسط مجموعه‌های فازی ناهموار خوشه‌بندی می‌کنند اما در این مقاله، رکوردهای درخواست بر اساس در نظر گرفتن شباهت همه ویژگی‌های آنها گروه‌بندی می‌شوند و شناسایی نشست‌های اولیه، توسط خوشه‌بندی مجموعه‌های فازی ناهموار به دست می‌آید.

ساختار ادامه مقاله به شرح زیر است: در بخش دوم، کارهای گذشته مورد بررسی قرار گرفته است. جزئیات روش پیشنهادی در بخش سوم بیان شده است. بخش چهارم نتایج آزمایش‌ها، ارزیابی و مقایسه آنها را نشان می‌دهد و در بخش پنجم، نتیجه‌گیری و کارهای آینده ذکر شده است.

۲- کارهای گذشته

در گذشته تشخیص روبات‌های وب توسط تکنیک‌های مختلفی انجام شده است. یکی از آنها تکنیک‌های مبتنی بر یادگیری تحلیلی[7] هستند. پژوهش‌های مبتنی بر این تکنیک ابتدا یک مرحله پیش‌پردازش شامل شناسایی نشست‌ها، استخراج ویژگی و برچسب‌گذاری نشست‌ها روی فایل ثبت وقایع انجام می‌دهند و سپس توسط الگوریتم‌های یادگیری به ارائه مدل تشخیص‌دهنده بازدیدکنندگان وب می‌پردازند.

به عنوان مثال، [۴] یک روش مدل‌سازی احتمالی توسط شبکه بیزین برای تشخیص روبات‌های وب ارائه داده است. ایده نویسندگان استفاده از حد آستانه پویا به عنوان زمان انقضا برای شناسایی نشست‌های بازدیدکنندگان بود. آنها ابتدا نشست‌ها را با مقایسه آدرس پروتکل اینترنتی و نام عامل کاربری درخواست‌ها گروه‌بندی کردند و سپس برای شکستن این گروه‌ها به زیرگروه‌ها از حد آستانه پویا استفاده کردند به طوری که متناسب با طول نشست‌ها حد آستانه نیز تغییر می‌کرد. در ادامه آنها شش ویژگی برای هر نشست، استخراج و نشست‌ها را توسط یک روش مکاشفه‌ای چهار مرحله‌ای برچسب‌گذاری کردند. در پژوهش دیگری، نویسندگان دو ویژگی جدید استخراج کردند که نشان‌دهنده رفتارهای خزنده‌های وب باشد. این دو ویژگی عبارتند از انحراف معیار عمق صفحات درخواست‌شده و درصد درخواست‌های Http پی در پی و پشت سر هم. آنها نشست‌ها را توسط مقایسه آدرس پروتکل اینترنتی و نام عامل کاربری و شکستن این گروه‌ها به زیرگروه‌ها توسط حد آستانه رایج ۳۰ دقیقه‌ای تعیین کردند و سپس ۹ ویژگی استخراج کردند؛ در ادامه آنها به ارزیابی این ویژگی‌ها توسط روش‌های طبقه‌بندی پرداختند [۵]. نویسندگان در [۲۶] از روش‌های طبقه‌بندی مبتنی بر ترکیب طبقه‌بندها[8] برای تشخیص روبات‌های وب استفاده کرده‌اند. آنها برای تعیین نشست‌های کاربران، ابتدا آنها را بر اساس نام عامل کاربری، گروه‌بندی و سپس از زمان انقضا برای شکستن این گروه‌ها استفاده کردند به طوری که متناسب با طول نشست، مقدار آن از ۲۵/۵ دقیقه تا ۲۴ ساعت متفاوت است. مرجع [۲۷] از یک روش نیمه‌نظارتی توسط ماشین بردار پشتیبان برای تشخیص روبات‌های وب استفاده کرده تا بتواند مزایای طبقه‌بندی نظارتی و غیر نظارتی را با هم داشته باشد. مرجع [۲۸] نیز یک روش طبقه‌بندی روبات‌های اینترنت با استفاده از شبکه بیزین ارائه داده که در آن از دو مرحله برای برچسب‌گذاری نشست‌های شناسایی‌شده استفاده کرده است. مرجع [۲۹] به روش دیگری به تشخیص روبات‌های وب پرداخته است. نویسندگان ابتدا یک وب سایت شامل مقالات و اخبار را

1. Fuzzy Rough Set Theory
2. Jaro-Winkler Distance
3. Session Identification
4. Feature Extraction
5. Heuristic
6. Session Labeling
7. Analytical Learning
8. Ensemble-Based Learner

بارگذاری و یک قالب مشخص برای ثبت درخواست‌ها به سرویس‌دهنده تعریف کردند. سپس در یک بازه خاص به جمع‌آوری این داده‌ها پرداخته و آنها را پردازش کردند. آنها برای شناسایی نشست‌های بازدیدکنندگان از بازه زمانی چهارساعته استفاده کردند. در ادامه توسط روش‌های طبقه‌بندی ماشین بردار پشتیبان (SVM)[1] و درخت تصمیم[2] C۴.۵ به طبقه‌بندی روبات‌های مخرب وب پرداختند. مرجع [۲۳] بعد از شناسایی نشست‌ها و استخراج ویژگی‌هایی برای هر نشست، با استفاده از تئوری مجموعه‌های فازی ناهموار، ویژگی‌هایی را انتخاب کرده است که بیشتر متمایزکننده رفتارهای کاربران وب است. سپس با استفاده از این ویژگی‌ها و مدل SOM نشست‌های روبات وب و کاربران انسانی را تعیین کرده است.

همچنین در حوزه طبقه‌بندی غیر نظارتی (خوشه‌بندی)، [۶] از دو الگوریتم شبکه عصبی SOM و ART۲ و همچنین از یک تحلیل‌کننده مبتنی بر جاوا برای پیش‌پردازش فایل ثبت وقایع سرویس‌دهنده و تعیین نشست‌ها استفاده کرده است. مرجع [۱۵] نیز با استفاده از روش خوشه‌بندی مبتنی بر چگالی و ارائه دو ویژگی جدید برای روبات‌های وب به تشخیص نوع بازدیدکننده‌های وب پرداخته است.

۳- روش پیشنهادی

روش پیشنهادی مبتنی بر تکنیک‌های یادگیری تحلیلی است. در این روش، چهار مرحله ارائه شده است که در ادامه به جزئیات هر مرحله پرداخته می‌شود.

۳-۱ شناسایی نشست‌ها توسط خوشه‌بندی مجموعه‌های فازی ناهموار

در مرحله شناسایی نشست‌ها، دنباله‌ای از درخواست‌های صادرشده از طرف یک بازیدکننده خاص شناسایی می‌شود. هرچه این شناسایی بهتر انجام شود، الگوهای (ویژگی‌های) استخراج‌شده بهتر ارائه می‌شوند و در نتیجه تشخیص نوع بازدیدکننده دقیق‌تر انجام می‌شود.

در این مقاله شناسایی نشست‌ها توسط خوشه‌بندی مبتنی بر نظریه مجموعه‌های فازی ناهموار انجام شده است. بر اساس این نظریه، مسأله تشخیص روبات‌های وب شامل یک سیستم اطلاعاتی[3] است که به صورت $IS=(U,A)$ نمایش داده می‌شود. در این سیستم، U[4] مجموعه محدودی از نمونه‌های مورد بحث را مشخص می‌کند و A[5] مجموعه‌ای از صفات را مشخص می‌کند که توصیف‌کننده نمونه‌های مجموعه جهانی می‌باشد. این صفات به دو دسته صفات شرطی[6] و صفت (صفات) تصمیم[7] تقسیم می‌شوند که در (۱) نشان داده شده‌اند

$$A \subseteq \{A \cup D\} \text{ such that } A \notin D \qquad (۱)$$

رابطه تشخیص‌ناپذیری[8] که یک رابطه هم‌ارزی است به صورت (۲) تعریف می‌شود

$$R_{ind} = \{(x,y) \,|\, \forall a \in A : a(x) = a(y)\} \qquad (۲)$$

1. Support Vector Machine
2. Decision Tree
3. Information System
4. Universal Set
5. Attributes
6. Condition Attributes
7. Decision Attribute
8. Indiscernibility Relation

این رابطه، اعضای مجموعه جهانی U را به مجموعه‌ای از کلاس‌های هم‌ارزی[9] تقسیم می‌کند که می‌تواند بیانگر گروه‌های اولیه باشد

$$\forall x \in U : [X]_{R_{ind}} = \{y \in U \,|\, \forall a \in A : a(x) = a(y)\} \qquad (۳)$$

در (۳)، x و y شامل رکوردهای درخواست بازدیدکنندگان وب می‌باشد.

کلاس‌های هم‌ارزی برای تعیین مفاهیم تقریب پایین[10] و تقریب بالا[11] که دو مفهوم پایه‌ای در نظریه مجموعه‌های ناهموار هستند، استفاده می‌شود. این مفاهیم به ترتیب در (۴) و (۵) نشان داده شده‌اند

$$\underline{A}_{R_{ind}} = \{X_i \in U \,|\, [X_i]_{R_{ind}} \subseteq A\} \qquad (۴)$$

$$\overline{A}_{R_{ind}} = \{X_i \in U \,|\, [X_i]_{R_{ind}} \cap A \neq \varnothing\} \qquad (۵)$$

در این سیستم، A به عنوان فضای تقریب[12] در نظر گرفته می‌شود که در مورد نمونه‌های متعلق به آن فضا مطمئن نیستیم، به همین دلیل از $\underline{A}_{R_{ind}}$ و $\overline{A}_{R_{ind}}$ که به ترتیب بیانگر تقریب پایین (شامل مجموعه عناصری که بدون شک متعلق به فضای تقریب هستند) و تقریب بالا (شامل مجموعه عناصری که احتمالاً متعلق به فضای تقریب هستند) استفاده می‌شود.

یکی از مشکلات نظریه مجموعه‌های ناهموار، کاربرد آن برای داده‌های گسسته و همچنین بیان گسسته مجموعه‌های تقریب بالا و پایین بود. برای رفع این مشکلات ایده فازی‌کردن این نظریه مطرح شد که Dubois و Prade یکی از اولین کسانی بودند که این ایده را مطرح کردند [۳۰]. در نظریه مجموعه‌های فازی ناهموار، $R(x,y)$ میزان مشابهت فازی دو نمونه x و y در فضای U^2 را بیان می‌کند و همچنین نظریه مجموعه‌های فازی ناهموار به مدل implicator/t-norm محدود شده است و فضاهای تقریب پایین و بالا به ترتیب به صورت (۶) و (۷) بیان می‌شود

$$\underline{R}(A)(x) = \inf_{y \in X} I(R(x,y), A(y)) \qquad (۶)$$

$$\overline{R}(A)(x) = \sup_{y \in X} T(R(x,y), A(y)) \qquad (۷)$$

که در آن I بیانگر عملگر Implicator و T بیانگر عملگر t-norm است [۳۰].

در این مقاله از $R(x,y)$ (رابطه بین دو نمونه) برای گروه‌بندی درخواست‌ها استفاده گردیده و از آنجایی که ویژگی‌های رکوردهای درخواست، اکثراً مقادیر رشته‌ای هستند از معیارهای شباهت رشته‌ای برای تعیین این رابطه استفاده شده است. این معیارها برای سنجش شباهت دو رشته به کار می‌روند که به سه دسته کلی معیارهای مبتنی بر تشابه لغوی (مقایسه ویژگی‌های رشته‌ای مانند نام و توضیحات)، معیارهای مبتنی بر تشابه ساختاری (شباهت دو موجودیت بر اساس ساختار آنها مانند آنتولوژی‌ها) و معیارهای مبتنی بر شباهت معنایی (شباهت موجودیت‌ها بر اساس نزدیکی معنای آنها مانند موتورهای جستجوی گوگل و یاهو) تقسیم می‌شوند [۳۱] و [۳۲].

لازم به ذکر است در این مقاله از معیارهای مبتنی بر تشابه لغوی استفاده شده است. معیارهای شباهت مختلفی در این دسته وجود دارند. با مطالعه معیارهای موجود در این زمینه و در نظر گرفتن رکوردهای

9. Equivalence Classes
10. Lower Approximation
11. Upper Approximation
12. Approximation Space

۱) **ورودی:** سیستم اطلاعاتی شامل (U, A) (فایل ثبت وقایع سرویس‌دهنده وب)
۲) به ازای تمام رکوردهای درخواست
۳) ایجاد کلاس هم‌ارزی شماره i
۴) به ازای تمام رکوردهای درخواست منهای رکورد شماره i
۵) اگر رکورد شماره j تاکنون گروه‌بندی نشده است:
۶) محاسبه شباهت کلیه فیلدهای دو رکورد درخواست طبق (۹)
۷) محاسبه شباهت نهایی دو رکورد درخواست با استفاده از محاسبه میانگین حسابی بردار شباهت
۸) اگر شباهت نهایی بیشتر از ۰٫۵شد، آن گاه رکورد شماره j را به کلاس هم‌ارزی شماره i اضافه کن.
۹) **خروجی:** کلاس‌های هم‌ارزی (گروه‌های اولیه نشست‌ها)

شکل ۱: الگوریتم شناسایی نشست‌های بازدیدکنندگان با استفاده از خوشه‌بندی مجموعه‌های فازی ناهموار و فاصله جارو- وینکلر.

درخواست مجموعه داده به این نتیجه رسیدیم که اندیس جاکارد، فاصله لونشتاین (فاصله ویرایشی)، فاصله جارو- وینکلر و معیار ضرایب مشابهت سورنسن (ضرایب تاس) می‌تواند شباهت دو رکورد درخواست را بهتر نشان دهند. این معیارها آزمایش شدند و با بررسی نتایج آنها مشخص شد معیار جارو- وینکلر بهتر می‌تواند شباهت دو رکورد درخواست را نشان دهد و بنابراین معیار شباهت (فاصله) جارو- وینکلر برای ارائه روش پیشنهادی استفاده شد. در علوم کامپیوتر و آمار، فاصله جارو- وینکلر معیاری برای اندازه‌گیری شباهت بین دو رشته است که نوعی از فاصله ویرایشی است و فاصله جارو- وینکلر کمتر بین دو رشته، شباهت بیشتر آن دو را نشان می‌دهد [۳۳]. این فاصله توسط (۸) محاسبه می‌شود

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{|m|}\right) & \text{otherwise} \end{cases} \quad (8)$$

که در آن m تعداد کاراکترهای تطبیق (مشترک بین دو رشته) و t نصف تعداد انتقالات برای تبدیل دو رشته را مشخص می‌کند.

بر اساس مطالب فوق در این مقاله شناسایی نشست‌ها با استفاده از رابطه شباهت درخواست‌ها انجام گردیده به طوری که برای هر دو درخواست، ابتدا شباهت ویژگی‌های متناظر آنها توسط معیار جارو- وینکلر محاسبه شده است. سپس میانگین حسابی این بردار شباهت، محاسبه می‌شود. در صورتی که این میانگین شباهت‌ها بیشتر از یک حد آستانه تعریف‌شده باشد این دو درخواست در یک نشست اولیه قرار می‌گیرند. در این مقاله از حد آستانه ۰٫۵ برای شباهت دو نمونه استفاده شده است. شکل ۱ الگوریتم شناسایی نشست‌ها را نشان می‌دهد.

لازم به ذکر است در ادامه برای شکستن گروه‌های ایجادشده به زیرگروه‌ها از روش زمان انقضای یک نشست استفاده شده که مشابه کارهای گذشته، این زمان معادل ۳۰ دقیقه در نظر گرفته شده است.

۳-۲ استخراج ویژگی

در مرحله استخراج ویژگی برای هر نشست تعدادی ویژگی جدید عددی که متمایزکننده رفتارهای بازدیدکنندگان وب است استخراج می‌شود. بیشتر این ویژگی‌ها در مقالات مختلف با هم همپوشانی دارند اما بسته به نظر نویسندگان این حوزه، ممکن است ویژگی‌های جدیدی استخراج شود. در این مقاله بر اساس مطالعه پژوهش‌های مرتبط [۴] تا [۶] و [۱۵]، تعداد ۱۰ ویژگی عددی که بیشتر متمایزکننده رفتارهای بازدیدکنندگان وب است استخراج شده است. این ویژگی‌ها عبارتند از تعداد کلیک‌ها (تعداد درخواست‌های یک فایل Http)، درصد درخواست‌های فایل‌های Pdf/Ps، درصد پاسخ خطای ۴xx، درخواست فایل Robots.Txt، درصد درخواست‌های تصویر، نرخ درخواست صفحات اینترنتی به تصویر، درصد درخواست‌های Http ارسال‌شده توسط روش درخواست Head، درصد درخواست‌های ارسال‌شده با ویژگی ارجاع تخصیص داده نشده، درصد درخواست فایل‌های Css و درصد پاسخ ۳xx.

۳-۳ برچسب‌گذاری نشست‌ها توسط روش مکاشفه‌ای

در مرحله برچسب‌گذاری نشست‌ها مشخص می‌شود نشست شناسایی‌شده از طرف کدام نوع بازدیدکننده وب درخواست شده است. هرچه این مرحله با صحت بیشتری انجام شود الگوی به دست آمده در مراحل بعدی با صحت بیشتری روبات‌های وب را تشخیص می‌دهد.

در این مرحله از یک روش برچسب‌گذاری مکاشفه‌ای برای تخصیص برچسب‌های اولیه هر نشست استفاده شده است. این روش سه ویژگی زیر را در رکورد درخواست بررسی می‌کند:

- درخواست فایل robots.txt: از آنجایی که تنها روبات‌های وب قادر به درخواست این فایل هستند، بررسی این ویژگی می‌تواند نوع بازدیدکننده را مشخص کند [۴].
- روش ارسال درخواست: روبات‌های وب بیشتر درخواست‌های خود را توسط روش head ارسال می‌کنند [۵].
- بررسی فیلد ارجاع[1]: فیلد ارجاع، آدرس مسیر جاری که درخواست از از آنجا ارسال شده را نشان می‌دهد. اکثر روبات‌های وب این فیلد را بدون مقدار ارسال می‌کنند [۴].

بنابراین در این مرحله ابتدا برای رکوردهای درخواست هر نشست، مقادیر سه ویژگی فوق بررسی و تعداد درخواست‌های روبات وب و کاربر انسانی محاسبه می‌شود. سپس نشست مربوط با تعداد برچسب‌های رایج آن برچسب‌گذاری می‌شود. در صورتی که برچسب‌های رایج یک نشست متمایز نباشد به صورت نشست ناشناخته برچسب‌گذاری می‌شود.

۳-۴ بهبود برچسب نشست‌ها توسط طبقه‌بندی مجموعه‌های فازی ناهموار

در این قسمت برای بهبود و افزایش اطمینان برچسب‌ها از مجموعه‌های فازی ناهموار استفاده شده است. این نظریه از مفهوم فضای تقریب پایین نمونه‌ها برای تصمیم‌گیری استفاده می‌کند که در (۹) نشان داده شده است

$$\underline{R}(A)(x) = \inf_{y \in X} I(R(x, y), A(y)) \quad (9)$$

همان طور که قبلاً بیان شد، I معادل عملگر Implicator می‌باشد که یک عملگر فازی است و کلاس‌های مختلفی از آن وجود دارد که در این

1. Referrer Field

۱) **ورودی:** سیستم تصمیم‌گیری شامل (U, A, AS) (مجموعه داده‌ای شامل نشست‌ها، ویژگی‌ها و برچسب نشست‌ها)
۲) ایجاد کلاس‌های (فضای تقریب) روبات وب و کاربر انسانی (بر اساس [۳۴])
۳) به ازای تمام رکوردهای درخواست شماره i منهای درخواست‌هایی که قطعاً عضو یکی از کلاس‌های مختلف است
۴) به ازای تمام رکوردهای درخواست شماره j منهای درخواست‌هایی که قطعاً عضو یکی از کلاس‌های مختلف است
۵) محاسبه رابطه دو رکورد درخواست شماره i و j طبق (۹)
۶) محاسبه عملگر implicator برای یک رکورد درخواست طبق (۱۴) برای هر دو کلاس
۷) محاسبه فضای تقریب پایین رکورد درخواست شماره i به هر دو کلاس طبق (۱۳)
۸) تعیین برچسب نهایی رکورد درخواست شماره i با مقایسه مقادیر فضاهای تقریب نسبت به هر دو کلاس (مقدار بزرگ‌تر)
۹) تعیین برچسب نهایی یک نشست با برچسب‌های رایج آن
۱۰) **خروجی:** برچسب‌های بهبودیافته نشست‌ها

شکل ۲: الگوریتم بهبود برچسب نشست‌ها توسط طبقه‌بندی مجموعه‌های فازی ناهموار.

جدول ۱: شناسایی نشست‌ها توسط خوشه‌بندی مجموعه‌های فازی ناهموار.

تعداد رکوردهای درخواست	تعداد نشست‌های شناسایی‌شده	زمان صرف‌شده برای گروه‌بندی اولیه درخواست‌ها (ثانیه)	تعداد نهایی نشست‌ها بعد از اعمال روش زمان انقضا
۲۰۰۰۰	۱۰۸۱	۱۴۲۶۳	۱۷۱۰

مقاله از کلاس S-Implicator و در آن از عملگر Kleene-Dienes استفاده شده است [۳۱]. رابطه (۱۰) این عملگر را نشان می‌دهد

$$I_{KD}(x, y) = \max\{1 - x, y\} \qquad (۱۰)$$

در (۱۰)، x بیانگر رابطه شباهت دو نمونه است که در مرحله شناسایی نشست‌ها محاسبه شده و y بیانگر تعلق یک نمونه به فضای تقریب (روبات وب و کاربر انسانی) است که در مرحله برچسب‌گذاری مکاشفه‌ای محاسبه شده است. در نهایت میزان تعلق همه نشست‌ها به فضای روبات وب و کاربر انسانی توسط (۹) (تقریب پایین نمونه‌ها) به دست می‌آید و سپس با بررسی آن در مورد برچسب نهایی نشست تصمیم‌گیری می‌شود. شکل ۲ الگوریتم بهبود برچسب‌ها را نشان می‌دهد.

لازم به ذکر است برای ساختن فضاهای تقریب مجموعه‌های فازی ناهموار از برچسب‌های مطمئن استفاده شده که در خط شماره ۲ الگوریتم شکل ۲ این فضاها ساخته می‌شود. به طور کلی در این مرحله رکوردهای درخواست برچسب‌گذاری می‌شود و سپس بر اساس آنها برچسب نهایی نشست مشخص می‌شود.

۴- نتایج آزمایش‌ها

در این بخش نتایج آزمایش‌های انجام‌شده ارائه و ارزیابی گردیده است. این آزمایش‌ها در محیط نرم‌افزار متلب R۲۰۱۳a بر روی پردازنده اینتل پنج هسته‌ای با سرعت ۲٫۶۹ گیگاهرتز و حافظه اصلی ۴ گیگابایت انجام گردیده و برای انجام آزمایشات از روش اعتبارسنجی متقاطع ۱۰تایی[1] استفاده شده است. این تکنیک برای تخمین کارایی یک مدل پیش‌بینی کننده استفاده می‌شود [۳۵] و به دلیل سادگی (ظاهری) و گستردگی جهانی [۳۶] در اکثر پژوهش‌های معتبر با وجود تعداد نمونه‌های زیاد مجموعه داده از این روش برای انجام آزمایش‌ها استفاده شده است [۲۲]، [۲۳]، [۳۷] و [۳۸]. در این مقاله، آزمایش‌ها روی فایل ثبت وقایع سرویس‌دهنده وب دانشگاه بین‌المللی امام رضا (ع) با آدرس دامنه www.imamreza.ac.ir انجام شده است. این فایل دارای فرمت ترکیبی است که حاوی ۲۰۰۰۰ رکورد درخواست است و در بازه ۱۵ April ۲۰۱۵ تا ۲۲ August ۲۰۱۵ ثبت شده است. برای انجام آموزش و طبقه‌بندی از الگوریتم‌های رایج طبقه‌بندی شامل ماشین بردار پشتیبان، درخت تصمیم و k نزدیک‌ترین همسایه[2] با پارامتر $k = ۳$ استفاده شده و همچنین برای نشان‌دادن برتری روش پیشنهادی، نتایج آزمایش‌ها با روش‌های مطرح در این زمینه مقایسه شده است.

اولین مرحله پیش‌پردازش، شناسایی نشست‌های بازدیدکنندگان وب است که یکی از ایده‌های اصلی مقاله می‌باشد. در این مرحله، رکوردهای درخواست توسط مجموعه‌های فازی ناهموار و بر اساس رابطه شباهت (فاصله) جارو- وینکلر خوشه‌بندی شده‌اند و جدول ۱ نتایج حاصل را نشان می‌دهد.

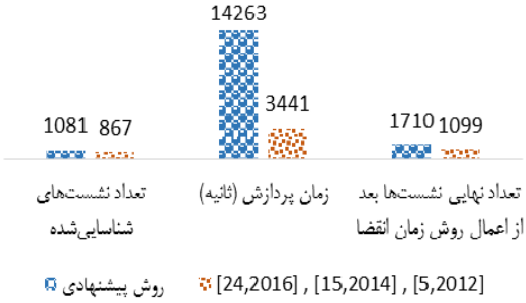
همان طور که در جدول ۱ آمده است زمان پردازش نسبتاً بالاست و به این خاطر است که رابطه شباهت تمام رکوردهای درخواست با هم محاسبه می‌شود. برای ارزیابی این روش، نتایج حاصل با نتایج روش رایج شناسایی نشست‌ها که فقط از دو ویژگی آدرس پروتکل اینترنتی و نام عامل کاربری درخواست‌ها برای گروه‌بندی استفاده می‌کند و در اکثر مقالات این حوزه استفاده می‌شود، مقایسه شده است. شکل ۳ این مقایسه را نشان می‌دهد.

همان طور که در شکل ۳ دیده می‌شود در روش رایج شناسایی نشست‌ها (روش استفاده‌شده در مقالات ذکرشده)، طول نشست‌ها بزرگ‌تر و تعداد نشست‌های کمتری شناسایی شده است به این خاطر که همه ویژگی‌ها مقایسه نمی‌شود. به عنوان مثال ویژگی تاریخ درخواست در این روش اهمیتی ندارد و بررسی مجموعه داده نشان داد این روش درخواست‌های روزهای مختلف را به عنوان یک نشست گروه‌بندی می‌کند. اما در روش پیشنهادی همه ویژگی‌های متناظر درخواست‌ها با هم مقایسه شده و شباهت نهایی دو درخواست، توسط میانگین شباهت همه ویژگی‌ها به دست آمده است. بنابراین نشست‌ها به طور دقیق‌تر و با صحت بیشتری شناسایی می‌شود.

بعد از استخراج ۱۰ ویژگی عددی برای هر نشست شناسایی‌شده، در مرحله بعد برچسب‌های اولیه نشست‌ها توسط روش مکاشفه‌ای تخصیص داده می‌شود. به منظور ارزیابی بهتر این روش برچسب‌گذاری، چهار آزمایش مختلف در نظر گرفته شده که این آزمایش‌ها در جدول ۲ توضیح داده شده‌اند.

1. 10-Fold Cross Validation

2. k-Nearest Neighbors (KNN)



شکل ۳: نمودار مقایسه روش پیشنهادی و روش رایج شناسایی نشست‌ها.

جدول ۲: آزمایش‌های تعریف‌شده برای ارزیابی روش پیشنهادی.

نام آزمایش	روش شناسایی نشست‌ها	روش برچسب‌گذاری
آزمایش اول	روش رایج	روش رایج
آزمایش دوم	روش پیشنهادی شناسایی نشست‌ها	روش رایج
آزمایش سوم	روش رایج	روش پیشنهادی مکاشفه‌ای
آزمایش چهارم	روش پیشنهادی شناسایی نشست‌ها	روش پیشنهادی مکاشفه‌ای

جدول ۳: نتایج انجام مرحله برچسب‌گذاری مکاشفه‌ای.

مجموعه داده	تعداد کل نشست‌ها	تعداد نشست‌های روبات وب	تعداد نشست‌های کاربر انسانی	تعداد نشست‌های ناشناخته
آزمایش اول	۱۰۹۹	۲۷۰	۰	۸۲۹
آزمایش دوم	۱۷۱۰	۱۶۹	۰	۱۵۴۱
آزمایش سوم	۱۰۹۹	۴۲۷	۴۷۳	۱۹۹
آزمایش چهارم	۱۷۱۰	۴۵۳	۹۰۵	۳۵۲

جدول ۴: مقایسه روش پیشنهادی و روش رایج شناسایی نشست‌ها.

روش شناسایی نشست‌ها	روبات‌های وب شناخته‌شده قبلی	روبات‌های وب جدید
روش رایج (آزمایش اول و سوم)	۶۲٫۹۹	۳۷٫۰۱
روش پیشنهادی (آزمایش دوم و چهارم)	۳۶٫۸۶	۶۲٫۹۹

لازم به ذکر است روش برچسب‌گذاری رایج در جدول ۲، روشی است که در بیشتر مقالات حوزه تشخیص روبات‌های وب مانند [۴] تا [۶] انجام شده است. این روش، دو ویژگی آدرس پروتکل اینترنتی و نام عامل کاربری رکوردهای درخواست هر نشست را با اطلاعات ثبت‌شده در پایگاه داده‌های موجود در اینترنت مقایسه می‌کند و سپس برچسب رایج این رکوردهای درخواست به عنوان برچسب نهایی نشست در نظر گرفته می‌شود. همان طور که جدول ۲ نشان می‌دهد، آزمایش چهارم روش پیشنهادی این مقاله را نشان می‌دهد. در ادامه برای ارزیابی روش پیشنهادی از این آزمایش‌های تعریف‌شده استفاده می‌شود.

بعد از اعمال مرحله برچسب‌گذاری مکاشفه‌ای، نتایج جدول ۳ حاصل شده است.

آزمایش اول و دوم برای نشان‌دادن تأثیر برچسب‌گذاری توسط روش رایج انجام شده است و همان طور که نتایج جدول ۳ نشان می‌دهد برچسب‌گذاری رایج وابسته به پایگاه داده‌ای است که اطلاعات بازدیدکنندگان اینترنت را جمع‌آوری نموده است. در این مقاله از تعداد ۲۱۱۸ اطلاعات مربوط به مرورگرهای اینترنت و ۸۳۷ اطلاعات مربوط به روبات‌های وب استفاده شده است [۳۴]. همان طور که گفته شد این اطلاعات ثابت هستند و وجود آنها در فایل ثبت وقایع، تضمین‌شده نیست. جدول ۳ نیز این مطلب را نشان می‌دهد و تعداد نشست‌های کاربر انسانی در آزمایش اول و دوم، صفر شده است اما از آنجایی که روبات‌های شناسایی‌شده توسط این روش قبلاً نیز شناسایی شده‌اند، برچسب‌های تخصیص داده شده به نشست‌ها از اطمینان قابل قبولی برخوردار هستند.

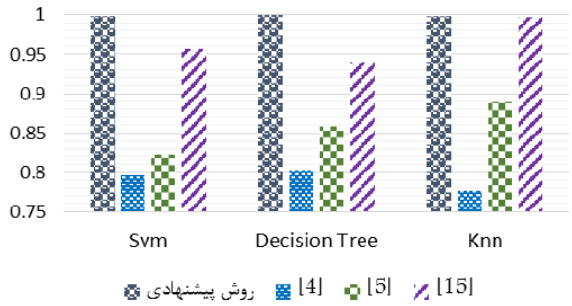
آزمایش سوم و چهارم برای نشان‌دادن کارایی روش برچسب‌گذاری پیشنهادی انجام شده است. بررسی نتایج این دو آزمایش در جدول ۳ نشان می‌دهد روش برچسب‌گذاری پیشنهادی برای نشست‌های شناسایی‌شده توسط هر دو روش، قادر به تشخیص روبات‌های وب و کاربران انسانی می‌باشد. نشست‌های ناشناخته در جدول ۳ نشست‌هایی هستند که درخواست‌های روبات‌های وب و کاربران انسانی در آنها با تعداد یکسانی توزیع شده‌اند.

با توجه به نتایج جدول ۳، میزان تأثیر روش پیشنهادی برچسب‌گذاری مکاشفه‌ای محاسبه شده و جدول ۴ این نتایج را نشان می‌دهد. با بررسی مقادیر این جدول می‌توان نتیجه گرفت زمانی که نشست‌ها توسط روش پیشنهادی شناسایی شده‌اند، روش برچسب‌گذاری پیشنهادی مکاشفه‌ای علاوه بر تشخیص روبات‌های وب قبلی بهتر می‌تواند نشست‌های روبات‌های وب را که اطلاعات آنها هنوز شناسایی نشده است تشخیص دهد. البته این واقعیت را باید در نظر گرفت که ممکن است روبات‌های وب قبلی از این سرویس‌دهنده بازدید نداشته‌اند.

بعد از انجام پیش‌پردازش فایل ثبت وقایع از الگوریتم‌های رایج طبقه‌بندی برای ایجاد مدل تشخیص‌دهنده و ارزیابی آن استفاده شده است. در مسأله تشخیص روبات‌های وب، کلاس‌ها (نشست‌های روبات وب و کاربر انسانی) دارای توزیع نامتوازن هستند. در این گونه مسایل معیار صحت به تنهایی نمی‌تواند کارایی طبقه‌بندی را نشان دهد و بنابراین از معیارهای دیگری نیز برای ارزیابی طبقه‌بندها استفاده می‌شود. جدول ۵ این معیارها را نشان می‌دهد.

برای مسأله تشخیص روبات‌های وب که یک مسأله دوکلاسه است، کلاس مثبت، نشست‌های روبات وب و کلاس منفی، نشست‌های انسان را توصیف می‌کند. همچنین T_p تعداد روبات‌های وبی که به درستی روبات وب شناسایی شده‌اند، F_n تعداد روبات‌هایی که به اشتباه انسان شناسایی شده‌اند، F_p تعداد انسان‌هایی که به اشتباه روبات وب در نظر گرفته شده‌اند و T_n تعداد انسان‌هایی که به درستی انسان تشخیص داده شده‌اند را مشخص می‌کند.

شکل ۴: نمودار افزایش معیار کارایی F بعد از بهبود برچسب نشست‌ها.



شکل ۵: نمودار نتایج مقایسه معیار F روش پیشنهادی و روش‌های مطرح.

جدول ۵: معیارهای ارزیابی کارایی طبقه‌بندها.

معیار	فرمول	کارایی
فراخوانی	$Recall = T_p/(T_p + F_n)$	طبقه‌بندی نمونه‌های مثبت (روبات وب)
دقت	$Precision = T_p/(T_p + F_p)$	دقت طبقه‌بندی (تشخیص درست روبات وب)
معیار F	$F = (۲\times Recall \times Precision)/(Recall + Precision)$	میانگین هارمونیک فراخوانی و دقت
صحت	$Accuracy = (T_p + T_n)/(T_p + F_p + T_n + F_n)$	نرخ پیش‌بینی درست طبقه‌بند

جدول ۶: نتایج معیارهای طبقه‌بندی.

طبقه‌بند	بعد از متوازن‌سازی برچسب‌ها			
	صحت	فراخوانی	دقت	معیار F
SVM	۹۹٫۸۰۰۴	۰٫۹۹۶۸	۰٫۹۹۹۰	۰٫۹۹۸۰
Decision Tree	۹۹٫۹۶۰۱	۱	۰٫۹۹۹۲	۰٫۹۹۹۶
KNN	۹۹٫۸۸۰۳	۰٫۹۹۸۴	۰٫۹۹۹۲	۰٫۹۹۸۸

جدول ۷: نتایج طبقه‌بندی نهایی بعد از تخصیص برچسب‌های بهبودیافته.

طبقه‌بند	بعد از بهبود برچسب نشست‌ها			
	صحت	فراخوانی	دقت	معیار F
SVM	۹۹٫۸۸۸۲	۰٫۹۹۸۶	۰٫۹۹۹۲	۰٫۹۹۸۹
Decision Tree	۹۹٫۹۶۰۲	۱	۰٫۹۹۹۲	۰٫۹۹۹۶
KNN	۹۹٫۹۴۴۲	۰٫۹۹۹۸	۰٫۹۹۹۱	۰٫۹۹۹۴

از آنجایی که توزیع برچسب‌های مجموعه داده به صورت نامتوازن[1] است نتیجه طبقه‌بندی ممکن است معتبر نباشد و بنابراین از روش متوازن‌سازی upsampling استفاده شده است. جدول ۶ نتایج طبقه‌بندی آزمایش ۴ (نشست‌های شناسایی‌شده و برچسب‌گذاری شده توسط روش پیشنهادی) را بعد از متوازن‌سازی برچسب نشست‌ها (کلاس‌ها) نشان می‌دهد. نتایج این جدول نشان می‌دهد با شناسایی بهتر نشست‌ها، برچسب‌گذاری مکاشفه‌ای و همچنین متوازن‌سازی برچسب‌ها می‌توان نتایج الگوریتم‌ها و معیارهای طبقه‌بندی و در نتیجه صحت تشخیص روبات‌های وب را افزایش داد.

در ادامه به منظور اطمینان از صحت برچسب‌های تخصیص داده شده، این برچسب‌ها توسط نظریه مجموعه‌های فازی ناهموار بهبود داده شده است. در این مرحله از تعداد ۱۲۵۷ نشستی که راجع به برچسب آنها مطمئن نبودیم، تعداد ۴ نشست جدید روبات وب شناسایی شد. این تعداد نشان می‌دهد که برچسب نشست‌ها با قطعیت بالایی شناسایی شده است به این علت که بهبود برچسب‌ها توسط مفهوم فضای تقریب پایین انجام شده و این مفهوم از حداقل شباهت و میزان تعلق درخواست‌ها به فضای تعریف‌شده (به شکل سخت‌گیرانه) برای تصمیم‌گیری نهایی استفاده می‌کند. بنابراین برچسب‌های تعیین‌شده از اطمینان قابل قبولی برخوردار هستند. بعد از این مرحله از تعداد ۱۷۱۰ نشست، ۴۵۷ نشست مربوط به روبات وب و ۱۲۵۳ نشست مربوط به کاربر انسانی است.

سپس مدل تشخیص‌دهنده بازدیدکنندگان وب بعد از تخصیص برچسب‌های بهبودیافته، ایجاد و ارزیابی شده است. جدول ۷ این نتایج را نشان داده می‌دهد. لازم به ذکر است این نتایج بعد از upsampling حاصل شده است.

به منظور نشان‌دادن کارایی مرحله بهبود برچسب نشست‌ها، شکل ۴ کارایی طبقه‌بند را از نظر معیار F نشان می‌دهد. با توجه به شکل می‌توان نتیجه گرفت که نظریه مجموعه‌های فازی ناهموار با در نظر گرفتن روابط ریاضی بین نمونه‌ها و میزان تعلق آنها به فضای خاص توانسته است در حد قابل قبولی تصمیم‌گیری مطمئن‌تری راجع به برچسب یک نمونه (نشست) انجام دهد.

همان طور که قبلاً ذکر شد، نوآوری مقاله در دو قسمت است: ۱) قسمت شناسایی نشست‌ها که نتایج جدول ۱، شکل ۳ و جدول ۴ کارایی روش پیشنهادی را نشان می‌دهد و ۲) بهبود برچسب نشست‌ها که شکل ۴، مشخص می‌کند بهبود برچسب‌ها می‌تواند معیارهای طبقه‌بندی را افزایش دهد.

همچنین برای نشان‌دادن کارایی روش پیشنهادی، این روش با روش‌های مطرح از نظر فراخوانی، صحت و معیار F، مقایسه شده است. از آنجایی که معیار F میانگین هارمونیک دو معیار فراخوانی و دقت است بهتر می‌تواند کارایی را نشان دهد. بنابراین نتایج معیار F در شکل ۵ نشان داده شده است.

۵- نتیجه‌گیری و کارهای آینده

در حوزه تشخیص روبات‌های وب، شناسایی نشست‌ها و برچسب‌گذاری آنها از اهمیت ویژه‌ای برخوردار است و هر چه این مراحل با دقت و صحت بیشتری انجام شود، تشخیص نهایی نوع بازدیدکنندگان نیز بهتر و دقیق‌تر انجام می‌شود. در این مقاله از مفاهیم خوشه‌بندی و طبقه‌بندی نظریه مجموعه‌های فازی ناهموار برای مسأله مبهم تشخیص روبات‌های وب استفاده شده است.

نظریه مجموعه‌های فازی ناهموار برای تصمیم‌گیری نهایی راجع به یک نمونه، رابطه آن را با تمام نمونه‌های مجموعه داده‌ها محاسبه می‌کند

1. Unbalanced

و بنابراین از دقت قابل قبولی برخوردار است اما در مسایلی که تعداد نمونه‌های مجموعه داده‌ها زیاد است، ممکن است زمان بیشتری مصرف کند که البته این زمان محاسبه به نوع رابطه تعریف‌شده نیز بستگی دارد. بنابراین یک مصالحه[1] بین معیار دقت و زمان پردازش وجود دارد و با به دست آوردن یکی از آنها ممکن است دیگری را از دست داد.

در ادامه به عنوان کارهای آینده در این زمینه می‌توان موارد زیر را مطرح کرد:

- بهبود مراحل مسأله تشخیص روبات‌های وب توسط نظریه مجموعه‌های فازی ناهموار، روش‌های مکاشفه‌ای و یا ترکیبی از هر دو.
- استفاده از معیارهای شباهت دیگر برای گروه‌بندی رکوردهای درخواست بازدیدکنندگان.
- استفاده از حد آستانه مناسب برای جداکردن گروه نشست‌ها به زیرگروه‌ها و همچنین حد آستانه مورد استفاده در تعیین شباهت دو نمونه درخواست.

مراجع

[1] D. Doran and S. S. Gokhale, "Web robot detection techniques: overview and limitations," *Data Min Knowl Disc*, vol. 22, no. 1-2, pp. 183-210, Jan. 2011.

[2] N. Algiriyage, S. Jayasena, G. Dias, A. Perera, and K. Dayananda, "Identification and characterization of crawlers through analysis of web logs," in *Proc. IEEE 8th Int. Conf.* on Industrial and Information Systems, *ICIIS'13*, pp. 150-155. Dec. 2013.

[3] J. Patel and H. Jethva, "Web crawling," *International J. of Innovations & Advancement in Computer Science*, vol. 4, pp. 228-235, May 2015.

[4] A. Stassopoulou and M. D. Dikaiakos, "Web robot detection: a probabilistic reasoning approach," *Computer Networks*, vol. 53, no. 3, pp. 265-278, Feb. 2009.

[5] D. Stevanovic, A. An, and N. Vlajic, "Feature evaluation for web crawler detection with data mining techniques," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8707-8717, Aug. 2012.

[6] D. Stevanovic, N. Vlajic, and A. An, "Detection of malicious and non-malicious website visitors using unsupervised," *Applied Soft Computing*, vol. 13, no. 1, pp. 698-708, Jan. 2013.

[7] D. Doran, *Detection, Classification, and Workload Analysis of Web Robots*, University of Connecticut, 2014.

[8] T. H. Sardar and Z. Ansari, "Detection and confirmation of web robot requests for cleaning the voluminous web log data," in *Proc. IEEE Int. Conf. on the IMpact of E-Technology on US, IMPETUS'14*pp. 13-19, Jan. 2014.

[9] Q. Bai, G. Xiong, Y. Zhao, and L. He, "Analysis and detection of bogus behavior in web crawler measurement," *Procedia Computer Science*, vol. 31, pp. 1084-1091, Dec. 2014.

[10] D. Doran, K. Morillo, and S. S. Gokhale, "A comparison of web robot and human requests," in *Proc. of the IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining, ACM*, pp. 1374-1380, Aug. 2013.

[11] M. D. Dikaiakosa, A. Stassopouloub, and L. Papageorgioua, "An investigation of web crawler behavior: characterization and metrics," *Computer Communications*, vol. 28, no. 8, pp. 880-897, May 2005.

[12] Z. Chu, S. Gianvecchio, A. Koehl, H. Wang, and S. Jajodia, "Blog or block: detecting blog bots through behavioral biometrics," *Computer Networks*, vol. 57, no. 3, pp. 634-646, Feb. 2013.

[13] D. Zhang, D. Zhang, and X. Liu, "A novel malicious web crawler detector: performance and evaluation," *IJCSI International J. of Computer Science Issues*, vol. 10, no. 1, pp. 121-126, Jan. 2013.

[14] I. Ghafir and V. Prenosil, "DNS traffic analysis for malicious domains detection," in *Proc. 2nd Int. Conf. on Signal Processing and Integrated Network,s SPIN'15*, pp. 613-918, Feb. 2015.

[15] M. Zabihi, M. V. Jahan, and J. Hamidzadeh, "A density based clustering approach to distinguish between web robot and human requests to a web server," *The ISC Int'l J. of Information Security*, vol. 6, no. 1, pp. 1-13, Jan. 2014.

[16] Z. Pawlak, "Rough sets," *International J. of Computer and Information Sciences*, vol. 11, no. 5, pp. 341-356, Oct. 1982.

[17] A. Anitha, "An efficient agglomerative clustering algorithm for web navigation pattern identification," *Circuits and Systems*, vol. 7, no. 9, pp. 2349-2356, Jul. 2016.

[18] R. Sadeghi and J. Hamidzadeh, "Automatic support vector data description," *Soft Computing*, 12 pp., 2016, DOI s00500-016-2317-5.

[19] K. Thangavel and R. Roselin, "Fuzzy-rough feature selection with Π-membership function for mammogram classification," *International J. of Computer Science Issues*, vol. 9, no. 4, pp. 361-370, May 2012.

[20] A. Zeng, T. Li, D. Liu, J. Zhang, and H. Chen, "A fuzzy rough set approach for incremental feature selection on hybrid information systems," *Fuzzy Sets and Systems*, vol. 258, pp. 39-60, Jan. 2015.

[21] N. Verbiest, *Fuzzy Rough and Evolutionary Approaches to Instance Selection*, Doctoral Dissertation, Ghent University, 2014.

[22] N. Verbiest, C. Cornelis, and F. Herrera, "FRPS: a fuzzy rough prototype selection method," *Pattern Recognition*, vol. 46, no. 10, pp. 2770-2782, Oct. 2013.

[23] J. Hamidzadeh, M. Zabihimayvan, and R. Sadeghi, "Detection of Web site visitors based on fuzzy rough sets," *Soft Computing*, 14 pp., 2016, DOI s00500-016-2476-4.

[24] D. U. Maheswari and A. Marimuthu, "An ensemble fuzzy rough set jaccard similarity measure based approach on user session clustering," *International J. of Computer Systems*, vol. 3, no. 4, pp. 330-334, Apr. 2016.

[25] T. V. Kumar and H. Guruprasad, "Clustering of web usage data using fuzzy tolerance rough set similarity and table filling algorithm," *Cancer Research and Oncology*, vol. 1, no. 3, pp. 143-152, Jun. 2013.

[26] D. S. Sisodia, S. Verma, and O. P. Vyas, "Agglomerative approach for identification and elimination of web robots from web server logs to extract knowledge about actual visitors," *J. of Data Analysis and Information Processing*, vol. 3, no. 1, pp. 1-10, Apr. 2015.

[27] W. Dong, *et al*., "Web robot detection with semi-supervised learning method," in *Proc. 3rd Int. Conf. on Material, Mechanical and Manufacturing Engineering, IC3ME'15*, pp. 2123-2128, 2015.

[28] G. Suchacka and M. Sobkow, "Detection of internet robots using a bayesian approach," in *Proc. 2nd IEEE Int. Conf. on Cybernetics, CYBCONF'15,* pp. 365-370, Jun. 2015.

[29] T. Grzinic, L. Mrsic, and J. Saban, *Lino-An Intelligent System for Detecting Malicious Web-Robots*, Intelligent Information and Database Systems, Springer International Publishing, pp. 559-568, 2015.

[30] A. M. Radzikowska and E. E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets and Systems*, vol. 126, no. 2, pp. 137-155, Mar. 2002.

[31] W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *Proc. American Association for Artificial Intelligence, IIWeb'03*, pp. 73-78, Acapulco, Mexico, 9-10 Aug. 2003.

[32] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International J. of Computer Applications*, vol. 68, no. 13, pp. 13-18, Jan. 2013.

[33] M. A. Jaro, "Probabilistic linkage of large public health data files," *Statistics in Medicine*, vol. 14, no. 5-7. pp. 491–498, Apr. 1995..

[34] List of User-Agents (Spiders, Robots, Browser), Retrieved from http://www.user-agents.org and www.UserAgentString.com, 2015.

[35] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, 2014.

[36] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40-79, 2010.

[37] S. Cifci, Y. Ekinci, G. Whyatt, A. Japutra, S. Molinillo, and H. Siala, "A cross validation of consumer-based brand equity models: driving customer equity in retail brands," *J. of Business Research*, vol. 69, no. 9, pp. 3740-3747, Sept. 2016.

[38] J. Hamidzadeh, R. Monsefi, and H. S. Yazdi, "IRAHC: instance reduction algorithm using hyperrectangle clustering," *Pattern Recognition*, vol. 48, no. 5, pp. 1878-1889, May 2015.

سمانه رحیمی در حال حاضر دانشجوی کارشناسی ارشد دانشگاه امام رضا (ع) است. ایشان مدرک کارشناسی خود را از دانشگاه صنعتی سجاد مشهد در سال ۱۳۹۱دریافت نموده‌اند. زمینه‌های علمی مورد علاقه نام برده عبارتند از: یادگیری ماشین، داده‌کاوی، تشخیص الگو.

جواد حمیدزاده در حال حاضر استادیار مهندسی کامپیوتر در دانشگاه صنعتی سجاد است. ایشان مدرک کارشناسی و کارشناسی ارشد خود را در رشته مهندسی کامپیوتر از دانشگاه صنعتی شریف به‌ترتیب در سال‌های ۱۳۷۴ و ۱۳۷۶ دریافت کرده‌اند. ایشان

1. Trade Off

مدرک دکترای خود را در رشته مهندسی کامپیوتر از دانشگاه فردوسی مشهد در سال ۱۳۹۱دریافت نموده‌اند. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: تشخیص الگو، یادگیری ماشین و محاسبات نرم.