الگویابی فعالیت‌های کاربران اینترنت با استفاده از روش‌های یادگیری ماشین

مجری

مریم میرزایی

همکاران

محمد شیری

عباس مرادی

پژوهشکده‌ی آمار

زمستان ۱۴۰۱

بسم الله الرحمن الرحیم

کد شناسه: RP-0103

عنوان: الگویابی فعالیت‌های کاربران اینترنت با استفاده از روش‌های یادگیری ماشین

مجری: مریم میرزایی

همکاران: محمد شیری، عباس مرادی

گروه پژوهشی: پردازش داده‌ها و اطلاع‌رسانی

تاریخ انتشار: زمستان ۱۴۰۱

نوبت انتشار: اول

طراح جلد: ساناز مهندسی

صفحه‌آرا: نجمه ناظریان

پژوهشکده‌ی آمار

تهران، خیابان دکتر فاطمی، خیابان باباطاهر، خیابان سرتیپ فکوری،
شماره‌ی ۱۴۵
۰۲۱ ۸۸۶۳۰۴۴۰ -۳
www.srtc.ac.ir

به نام خداوند جان و خرد

پیش‌گفتار

در عصر جدید، افراد بسیاری فعالیت‌های روزمره خود اعم از خرید، استفاده از خدمات آموزشی، فرهنگی، تراکنش‌های بانکی، مطالعه و ... را به صورت اینترنتی انجام می‌دهند. در این طرح علاقه‌مند به شناسایی الگوهای فعالیت‌های کاربران با تحلیل داده‌های گذران وقت در فضای مجازی و اینترنت هستیم. برای این منظور با توجه به ضرورت مدرن‌سازی نظام آماری کشورها و اهمیت استفاده از روش‌های جدید در آمار رسمی، روش‌های یادگیری ماشین مورد استفاده قرار گرفتند. یکی از روش‌های یادگیری ماشین خوشه‌بندی است. این یک روش یادگیری ناراهنماییده است، از این‌رو هیچ نظارتی برای الگوریتم ارائه نمی‌شود و با مجموعه داده بدون برچسب سروکار دارد. الگوریتم خوشه‌بندی تا حدودی شبیه به الگوریتم رده‌بندی است، اما تفاوت در نوع مجموعه داده‌ای است که استفاده می‌شود. به طور کلی می‌توان گفت هدف اصلی این طرح شناسایی الگوهای پنهان استفاده از اینترنت و فضای مجازی و همچنین بررسی تأثیرگذاری عوامل مختلف بر نوع استفاده از اینترنت و فضای مجازی است. در این طرح از آنجایی که با داده‌های آمیخته سروکار داریم استفاده از روش‌های تک بعدی که فقط برای خوشه‌بندی داده‌های عددی و یا خوشه‌بندی داده‌های رسته‌ای استفاده می‌شوند، مناسب نیست و از این رو از روش‌های خوشه‌بندی برای داده‌های نوع آمیخته استفاده می‌کنیم.

مطالعه حاضر به این صورت سازماندهی شده است؛ ابتدا مقدمات موضوع مورد بحث قرار می‌گیرد، سپس مفاهیم و روش‌های یادگیری ماشین مانند رده‌بندی و خوشه‌بندی مطرح می‌شوند و درنهایت روش‌های یادگیری ماشین برای تحلیل داده‌های گذران وقت مرکز آمار مورد استفاده قرار می‌گیرند.

گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی

پژوهشکده‌ی آمار

فهرست مطالب

فهرست جدول‌ها

فهرست شکل‌ها

۱

کلیات تحقیق

۱-۱- مقدمه

گذران وقت در فضای مجازی و اینترنت یکی از موضوع‌های مهمی است که امروزه در زندگی بشر جا باز کرده است و بسیاری از افراد وقت زیادی از زندگی روزمره خود را صرف فعالیت در فضای مجازی و اینترنت می‌کنند. حال ممکن است این فعالیت‌ها منجر به اتلاف وقت و یا صرفه‌جویی در وقت فرد شود. فعالیت‌هایی مانند خرید، آموزش، سرگرمی، دستیابی به اطلاعات عمومی، تماشا کردن برنامه‌ها یا گوش کردن به موسیقی، جستجو در شبکه‌های اجتماعی، دانلود و آپلود کردن اطلاعات، چک کردن ایمیل و ... می‌تواند فرد را ترغیب به استفاده از رایانه و اینترنت کند. از آنجایی که متغیرهای بسیاری مانند سن، جنسیت، تحصیلات، وضعیت تأهل، شغل، شهری یا روستایی بودن، تعداد اعضای خانوار و ... می‌توانند بر نوع استفاده افراد از اینترنت و فضای مجازی تأثیرگذار باشد، یکی از اهداف این طرح بررسی تأثیرگذاری این متغیرها بر نوع فعالیت افراد در اینترنت و فضای مجازی است. یکی دیگر از اهداف این پژوهش یافتن الگوهای استفاده افراد از اینترنت و فضای مجازی است. که برای دستیابی به این اهداف روش‌های یادگیری ماشین[1] مورد استفاده قرار می‌گیرد. برای این منظور سعی بر این است که تمرکز روی روش‌های نوین تحلیل داده از جمله روش‌های یادگیری ماشین باشد و بهترین روش برای دستیابی به اهداف ذکر شده مورد استفاده قرار گیرد. به طور کلی می‌توان گفت در طول تحقیق باید به سؤالات زیر پاسخ داده شود.

- چه عواملی بر نوع استفاده از اینترنت و فضای مجازی تأثیرگذار هستند؟
- الگوی استفاده از اینترنت و فضای مجازی به چه صورت است؟
- به طور کلی افراد با چه ویژگی مانند سن، تحصیلات، وضعیت تأهل، وضعیت اشتغال و ... از چه الگوی مصرفی اینترنت تبعیت می‌کنند؟

---

[1] Machine Learning

- افراد با گروه‌های سنی مختلف چه نوع فعالیت و چه مدت زمان از اینترنت استفاده می‌کنند؟
- کدام روش یادگیری ماشین برای تحلیل داده‌های اینترنت و فضای مجازی مطرح شده‌اند؟
- کدام روش برای استخراج الگوی استفاده از فضای مجازی و اینترنت مفید است؟

در این فصل ابتدا پیشینه‌ای از پژوهش‌های انجام شده در زمینه فضای مجازی و اینترنت و همچنین روش‌های یادگیری ماشین مطرح می‌شود سپس به مفاهیم کلی پژوهش از قبیل ضرورت و اهمیت موضوع، اهداف، روش تحقیق و تعاریف و مفاهیم ضروری پرداخته می‌شود.

۱-۲- پیشینه تحقیق

در سال‌های اخیر مطالعات بسیاری در رابطه با استفاده از اینترنت و فضای مجازی انجام شده است که در ادامه به برخی از این مطالعات اشاره شده است. حبیبی و بهنامی (۱۳۹۵) به بررسی جایگاه فضای مجازی و اینترنت و فضاهای شهری و مقایسه‌ی دوام و کیفیت حضور نوجوانان محله‌ی رجایی شهر در هر یک از این فضاها پرداخته‌اند. این پژوهش روی ۲۵۰ نوجوان صورت گرفته است. یافته‌ها نشان می‌دهد که روزانه ۸۸ درصد از نوجوانان در اینترنت و ۴۷ درصد از آنان در فضاهای شهری حضور می‌یابند که بین متغیرهای زمینه‌ای، پایگاه اجتماعی- اقتصادی، امنیت اجتماعی و هویت اجتماعی آنان با میزان حضورشان در این فضاها رابطه‌ی معناداری وجود دارد. اما بین میزان حضورشان در فضاي مجازی با متغیرهای جنسیت، امنیت جانی و امنیت مالی رابطه‌ی معناداری حاصل نشد. در مجموع ۵۹ درصد از نوجوانان، در ساعات عصر که بهترین بخش از اوقات فراغتشان برای حضور در فضاهای شهری تلقی می‌گردد، مشغول فعالیت در فضای مجازی و اینترنت هستند و طبق یافته‌های استنباطی، هرچه میزان حضور نوجوانان در فضای مجازی و اینترنت بیشتر میزان گرایش آن‌ها به حضور در فضاهای شهری کاسته می‌شود.

در تحقیقات گذشته طیف گسترده‌ای از عوامل مؤثر بر پذیرش و استفاده از رایانه مورد بررسی قرار گرفته است، مانند مطالعات دیویس و همکاران (۱۹۸۹)، برانچیو و ودرِب (۱۹۹۰)، ایگباریا و همکاران (۱۹۹۴)، تانگ (۱۹۹۹) و گفن و استراوب (۲۰۰۰). اتکینسون و کید (۱۹۹۷) عوامل مؤثر بر استفاده از اینترنت را مورد بررسی قرار دادند. در این مطالعه دریافتند که سرگرمی و کار و سودمندی بر استفاده از اینترنت تأثیرگذارند. تئو (۲۰۰۱) با ادامه کار اتکینسون و کید به بررسی ارتباط بین متغیرهای جمعیت شناختی و انگیزشی و نوع استفاده از اینترنت (فعالیت عمومی پیام رسانی، مرور، دانلود و خرید) پرداخت. با تداوم رشد سریع اینترنت، درک عوامل جمعیت شناختی و انگیزشی مرتبط با استفاده از اینترنت حیاتی است. از آنجایی که اینترنت دارای مزایایی از جمله تنوع در ارائه دانش، سرعت و راحتی دسترسی به اطلاعات است، جوانان امروزی برای کسب اطلاعات، استفاده از اینترنت را بر هر چیز دیگری ترجیح می‌دهند، بنابراین بررسی نیازهای آنها در آموزش و استفاده از رویکردهای نوین آموزشی امری ضروری است (بیلگیچ و همکاران، ۲۰۱۱). در همین راستا بیلگیچ و همکاران (۲۰۱۱) به بررسی و تعیین تأثیر عوامل جمعیت شناختی بر اهداف استفاده از اینترنت دانش‌آموزان دبیرستانی پرداخت. جامعه مورد مطالعه شامل دانش آموزان پایه‌های نهم تا دوازدهم دبیرستان‌های آناتولی، دبیرستان‌های علوم، دبیرستان‌های علوم اجتماعی، دبیرستان‌های ورزشی و دبیرستان‌های هنرهای زیبا در ترکیه بود.

ساوان و جایوسی (۲۰۲۰) با بکارگیری روش‌های یادگیری با نظارت یا راهنماییده[۲] و ناراهنماییده[۳] به رده‌بندی[۴] فعالیت‌های کاربران اینترنت می‌پردازند. در این مطالعه از الگوریتم ناراهنماییده k-میانگین[۵] جهت الگویابی و در نهایت از الگوریتم جنگل تصادفی[۶] برای به دست آوردن برچسب‌های فعالیت استفاده می‌کنند. کواچویچ و کاسلان (۲۰۲۰) به بررسی الگوهای استفاده از اینترنت بر اساس جنسیت به منظور برجسته کردن تفاوت‌های مربوط به شدت استفاده و نوع فعالیت‌های آنلاین پرداختند در این مطالعه برای تحلیل داده‌ها روش یادگیری عمیق، خوشه‌بندی k-میانگین و درخت تصمیم را مورد استفاده قرار دادند. این پژوهش جهت دستیابی به درک بهتری از نحوه ارتباط فعالیت‌های آنلاین با تفاوت‌های جنسیتی کمک می‌کند و تأیید می‌کند که بکارگیری روش‌های یادگیری عمیق[۷] می‌تواند به طور مؤثر این تفاوت‌ها را شناسایی کند. لاباین و همکاران (۲۰۲۰) اخیراً پژوهشی را انجام دادند که در آن برنامه‌ای برای رده‌بندی فعالیت‌های کاربر با استفاده از یادگیری راهنماییده و ناراهنماییده ارائه شده است، این برنامه از رفتار نمایش داده شده در شبکه استفاده می‌کند و فعالیت کاربر را با در نظر گرفتن تمام ترافیک تولید شده توسط کاربر در یک پنجره زمانی مشخص رده‌بندی می‌کند.

روش‌های بسیاری برای تحلیل داده‌ها وجود دارد، اما در عصر جدید با بالا رفتن حجم داده‌ها دیگر بکارگیری روش‌های سنتی پاسخگو نیست و بهتر است روش‌های جدید مورد استفاده قرار گیرد. در سال‌های اخیر روش‌های بسیاری تحت عنوان روش‌های یادگیری ماشین مطرح شده است که برای تحلیل داده‌ها با حجم‌های بالا نیز بسیار مناسب است. یادگیری ماشین زیر شاخه‌ای از هوش مصنوعی است و یکی از جنبه‌های ضروری تجارت و تحقیقات مدرن برای بسیاری از سازمان‌های امروزی است. الگوریتم‌های یادگیری ماشین به‌طور خودکار یک مدل ریاضی با استفاده از داده‌های نمونه- که به عنوان « مجموعه داده آموزشی[۸]» نیز شناخته می‌شود- می‌سازند تا بدون برنامه‌ریزی خاص تصمیم‌گیری کنند. در عین حال ساده‌ترین تعریف برای یادگیری ماشین، این است؛ چگونه ماشین‌ها از داده‌ها با کشف الگوهای آماری برای تصمیم‌گیری و انجام وظایف به تنهایی آموزش داده می‌شوند. الگوریتم‌های یادگیری ماشین عموماً می‌توانند راهنماییده، ناراهنماییده یا یادگیری‌های تقویتی[۹] باشند (کریستوفر، ۲۰۰۶، برادشاو و همکاران، ۲۰۱۳، ساموئل، ۱۹۸۸، سز و همکاران، ۲۰۱۷). یک جهت تحقیقاتی مهم در زمینه یادگیری ماشین، تحلیل خوشه‌ای است که متعلق به یادگیری ناراهنماییده است. تحلیل خوشه‌ای به عنوان ابزار مهم تحلیل داده می‌تواند اشیاء داده را با محاسبه عدم تشابه اشیاء داده بدون برچسب به زیرخوشه‌های مختلف تقسیم کند. در این مطالعه هدف دستیابی به این است که اشیاء داده در یک خوشه مشابه دارای عدم تشابه کمتر و اشیاء داده در خوشه‌های متفاوت دارای عدم تشابه بیش‌تری باشد (لام و همکاران، ۲۰۱۱).

---

[2] Supervised Learning
[3] Unsupervised Learning
[4] Classification
[5] KMean
[6] Random forest
[7] Deep learning
[8] training dataset
[9] Reinforcement learning