مجله‌ی بررسی‌های آمار رسمی ایران
سال۲۴، شماره‌ی ۱، بهار و تابستان ۱۳۹۲، صص ۱-۱۷

به کارگیری فن‌های داده‌کاوی برای بررسی ویژگی‌های کاربران اینترنت استان تهران

سحر صاحبی[†,*] و سپیده صاحبی[‡]

[†] مرکز آمار ایران
[‡] شرکت ملی نفتکش ایران

چکیده: داده‌کاوی ابزاری است که ما را قادر به بررسی حجم گسترده‌ای از داده‌ها، یافتن ارتباط آن‌ها، تحلیل روابط، پیش‌بینی نتایج و ایجاد مدل‌های کاربردی می‌سازد. از آنجا که گسترش اینترنت در زندگی امروزی به ابزاری برای رشد و توسعه‌ی کشورها تبدیل شده و میزان نفوذ آن در حوزه‌های مختلف شاخصی برای رشد و توسعه‌ی کشورها در نظر گرفته می‌شود، لذا می‌توان از فن‌های داده‌کاوی در تحلیل نفوذ اینترنت در زندگی اجتماعی استفاده نمود. سند چشم‌انداز بیست ساله، توسعه‌ی ارتباطات و زیرساخت‌های ارتباطی و فن‌آوری اطلاعات متناسب با پیشرفت‌های جهانی را هدف قرار داده است. از این رو دولت موظف به گسترش ارایه‌ی خدمات الکترونیکی است. این مقاله به مطالعه‌ی الگوهای رفتاری کاربران و ویژگی‌های آن‌ها پرداخته است و به‌علت این که روابط و پارامترهای بدست‌آمده بر حسب شرایط اجتماعی هر جامعه متفاوت خواهند بود، ضرورت انجام این مطالعه حائز اهمیت است. سپس داده‌های طرح آمارگیری از کاربران اینترنت در سطح استان تهران برای سال ۱۳۸۹ مورد بررسی قرار گرفته و با کمک فن‌های داده‌کاوی، کاربران خوشه‌بندی شده، و با استفاده از قواعد انجمنی روابطی میان ویژگی‌های مختلف کاربران بدست آمد که نتایج آن می‌تواند برای برنامه‌ریزی در خصوص گسترش استفاده از اینترنت در حوزه‌های مختلف مورد استفاده قرار گیرد.

واژگان کلیدی: داده‌کاوی؛ اینترنت؛ کاربر اینترنت؛ خوشه‌بندی؛ قواعد انجمنی.

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۹۱/۵/۳، پذیرش: ۱۳۹۳/۵/۱.

۱- مقدمه

داده‌کاوی، ابزاری است که به‌منظور تجزیه و تحلیل داده‌ها و خلاصه کردن آن‌ها به‌منظور دستیابی به اطلاعاتی مفید و قابل استفاده، بکار گرفته می‌شود. ابزارهای داده‌کاوی این امکان را برای محققان فراهم می‌سازد که حجم وسیع داده‌ها را از جنبه‌های مختلف بررسی و طبقه‌بندی کرده و روابط بدست آمده را مورد استفاده قرار دهند. با کمک فن‌های داده‌کاوی، می‌توان به روابط علی و معلولی بین داده‌ها دست یافت. در داده‌کاوی دو الگوی متفاوت برای یادگیری موجود است: یادگیری با ناظر[۱] و یادگیری بدون ناظر[۲]. در یادگیری با ناظر، ویژگی‌ها و پاسخ‌های یک مسئله‌ی خاص به سیستم یادگیری ارایه می‌شود و پس از آن سیستم قادر خواهد بود، برای مسائل مشابه که جوابشان مشخص نیست، پاسخ‌هایی را ارایه دهد. دو رویکرد شناخته‌شده برای این فرایند یادگیری، شبکه‌ی عصبی و درخت تصمیم است. در یادگیری بدون ناظر، پاسخ‌های مسئله مشخص نیستند و رفتار و پاسخ بهینه ارایه نمی‌شود. اما بر اساس تصمیم اتخاذشده، جریمه و پاداش در نظر گرفته می‌شود. خوشه بندی[۳] و قوانین انجمنی[۴] از این دسته هستند [۵].

اینترنت از پیوند تعداد بی‌شماری شبکه‌های ارتباطی کامپیوتری کوچک و بزرگ که حاوی اطلاعات متنوع است، تشکیل شده است. یک فرد متصل به شبکه اینترنت تنها مشاهده‌گر ومرورگر اینترنت نیست، بلکه جزیی از این شبکه بوده و می‌تواند با آن تبادل اطلاعات نماید. در دنیای امروز اینترنت، ابزاری برای رشد و توسعه‌ی کشورها تبدیل شده و تأثیر آن هر روز در ابعاد مختلف زندگی بیش‌تر می‌شود. بر اساس تعریف ارایه‌شده توسط مرکز آمار ایران، کاربر اینترنت فردی است که در یک دوره‌ی ۱۲ ماهه، در داخل کشور، با اتصال به شبکه اینترنت به هر مدت، شخصاً حداقل از یکی از خدمات اینترنتی استفاده کرده باشد. خدمات اینترنتی شامل دریافت اطلاعات، پست الکترونیکی، تعامل دولتی، بانک‌داری اینترنتی، آموزش اینترنتی، خرید و سفارش کالا، تلفن زدن اینترنتی، دانلود بازی، فیلم، عکس و غیره می‌شود [۱].

در سال ۲۰۰۶، اوکازاکی [۴] به بررسی الگوهای رفتاری کاربران اینترنت از طریق تلفن همراه در ژاپن پرداخته و با استفاده از رویکرد خوشه‌بندی، کاربران را با توجه به پارامترهایی مانند جنسیت، سن، وضع اشتغال، وضعیت تأهل و غیره به چهار خوشه تقسیم کرد.

طرح آمارگیری از کاربران اینترنت، برای اولین بار در دی ماه ۱۳۸۷ توسط مرکز آمار ایران اجرا شد. پس از آن این طرح در سال ۱۳۸۹ مجدداً به اجرا درآمد و اطلاعات کاربران را در دوازده ماه منتهی به زمان آمارگیری جمع‌آوری کرد. تا قبل از این آمارگیری‌ها، اطلاعی رسمی از ضریب نفوذ اینترنت وجود نداشت و نرخ‌های اعلام‌شده برای ضریب نفوذ در سال‌های قبل بر اساس براوردها و استفاده از فرمول‌های محاسباتی بوده است. یکی از موضوعات جالب در این حوزه، بررسی و مطالعه‌ی ویژگی‌های کاربران اینترنت است، که با توجه به مشخصات فردی، تحصیلی، وضعیت شغلی و غیره‌ی افراد متفاوت است و شناخت آن‌ها می‌تواند متخصصان را برای برنامه‌ریزی‌های آینده یاری رساند. با این وجود مطالعات اندکی در خصوص ویژگی‌های کاربران اینترنت و الگوهای رفتاری آن‌ها در ایران انجام شده است. در این مقاله، به بررسی ویژگی‌های کاربران اینترنت در سطح استان تهران – به‌عنوان استانی که بیش‌ترین ضریب نفوذ اینترنت را در کل کشور داشته است– بر اساس اطلاعات سال ۱۳۸۹ پرداخته‌ایم و کاربران در سطح استان تهران با رویکرد خوشه‌بندی به چهار خوشه تقسیم شده‌اند که افراد هر خوشه دارای ویژگی‌های رفتاری شبیه به یکدیگر می‌باشند. همچنین تلاش شده، تا برخی قوانین معنی‌دار موجود میان متغیرها از طریق قوانین انجمنی بیان شود. نتایج بدست آمده از این مطالعه می‌تواند برای متخصصان این حوزه، مراکز ارایه‌دهنده‌ی خدمات اینترنتی، بانک‌ها و مؤسسات مالی، مؤسسات آموزشی و غیره جالب توجه باشد.

این مقاله به‌صورت زیر سازمان‌دهی شده است: بخش دوم، به بیان نحوه‌ی نمونه‌گیری پرداخته است. متغیرهای مورد استفاده، در بخش سوم توضیح داده شده و سپس در بخش چهارم از رویکردهای داده‌کاوی برای بررسی داده‌ها استفاده شده است. نتایج بدست آمده، در بخش پنجم مورد تجزیه و تحلیل قرار گرفته‌اند و جمع‌بندی نهایی این مطالعه در بخش ششم ذکر شده است.

۲- شیوه‌ی نمونه‌گیری

چارچوب نمونه‌گیری مرحله‌ی اول برای خانوارهای معمولی ساکن و گروهی، فهرست حوزه‌های سرشماری عمومی نفوس و مسکن سال ۱۳۸۵ در کشور است. این فهرست علاوه بر اطلاعات جغرافیایی حوزه‌ها، تعداد خانوارهای معمولی ساکن و گروهی و جمعیت حوزه‌ها، شامل اطلاعات متغیرهای کمکی مورد نیاز برای طبقه‌بندی حوزه‌ها نیز

می‌باشد. چارچوب نمونه‌گیری مرحله‌ی دوم، برای خانوارهای معمولی ساکن و گروهی، فهرست خوشه‌ها در حوزه‌های نمونه است که بر اساس فهرست بلوک/آبادی سرشماری ۱۳۸۵ ساخته شده است. چارچوب نمونه‌گیری مرحله‌ی سوم، برای خانوارهای معمولی ساکن و گروهی در خوشه‌های نمونه بر اساس اطلاعات فرم فهرست‌برداری سرشماری ۱۳۸۵ است. نمونه‌های این طرح برای خانوارهای معمولی ساکن و گروهی طی سه مرحله و برای خانوارهای مؤسسه‌ای طی دو مرحله و با استفاده از روش نمونه‌گیری احتمالی انتخاب شده‌اند. برای نمونه‌گیری از خانوارهای معمولی ساکن و گروهی، در مرحله‌ی اول، ابتدا حوزه‌های هر استان بر اساس نقاط شهری و روستایی و نیز شهرستان مرکز استان و سایر شهرستان‌های استان، در چهار طبقه قرار گرفتند، سپس بر اساس پنج متغیر کمکی مناسب موجود در چارچوب که عبارت‌اند از نسبت سالمندان، دانشگاهیان، محصلان، سرپرستان مرد، خانوارهای برخوردار از رایانه، یک مؤلفه‌ی اصلی بدست آمده که در بیش‌تر استان‌ها، همبستگی بیش از ۸۰ درصد با متغیر نسبت خانوارهای استفاده‌کننده از اینترنت داشت. پس از آن، با استفاده از این مؤلفه‌ی اصلی و متغیر نسبت خانوارهای استفاده‌کننده از اینترنت و با کمک روش خوشه‌بندی فازی، حوزه‌ها در هر یک از طبقات مرحله‌ی قبل، به دو دسته تقسیم شدند. به این ترتیب، حوزه‌هایی که بر اساس متغیرهای خوشه‌بندی فازی مشابه یکدیگر هستند، در یک طبقه و سایر حوزه‌ها در طبقه‌ی دیگر قرار گرفتند. بر این اساس حوزه‌های هر استان، بجز استان قم که فقط یک شهرستان مرکز استان دارد، به هشت طبقه تقسیم شدند. سپس در هر طبقه‌ی هر استان، حوزه‌های نمونه به روش تصادفی ساده انتخاب شدند. در مرحله‌ی دوم نمونه‌گیری، درون هر یک از حوزه‌های انتخاب‌شده، خوشه‌های نمونه به روش تصادفی ساده انتخاب شدند. در این طرح هر خوشه عبارت است از یک بلوک/آبادی یا مجموعه‌ای از چند بلوک یا آبادی کوچک، به این ترتیب که چنان‌چه بلوک/آبادی شامل حداقل ۱۰ خانوار باشد، به تنهایی یک خانوار را تشکیل می‌دهد ولی اگر کم‌تر از ۱۰ خانوار داشته باشد، آن بلوک/آبادی با بلوک/آبادی مجاور ادغام شده و تشکیل یک خوشه می‌دهند. در مرحله‌ی سوم نمونه‌گیری، داخل هر یک از خوشه‌های نمونه، ده خانوار به روش تصادفی ساده انتخاب شدند که از بین آن‌ها پنج خانوار به‌عنوان نمونه‌ی اصلی و پنج خانوار به‌عنوان نمونه‌ی جایگزین در نظر گرفته شدند. برای این طرح ۵۳۷۲۴ خانوار نمونه، شامل ۵۲۹۸۵ خانوار معمولی و گروهی و ۷۳۹ خانوار مؤسسه‌ای انتخاب شده است.

۳- متغیرهای مورد استفاده

متغیرهای در نظر گرفته شده در طرح کاربران اینترنت عبارت بودند از، تعداد کاربران اینترنت به‌تفکیک گروه‌های جنسی و سنی، وضع سواد، دوره یا مدرک تحصیلی، وضع فعالیت، تعداد کاربران اینترنت به‌تفکیک نوع محل‌های دسترسی به اینترنت، تعداد کاربران اینترنت به‌تفکیک نوع دسترسی به اینترنت در منزل، متوسط دفعات دسترسی به اینترنت، متوسط مدت زمان استفاده از اینترنت، تعداد کاربران اینترنت به‌تفکیک نوع استفاده از اینترنت، تعداد کاربران اینترنت به‌تفکیک برخورداری خانوارها از رادیو، تلویزیون، تلفن ثابت، همراه یا رایانه در منزل.

در این مطالعه، تعدادی از متغیرهای فوق مورد بررسی قرار گرفتند، که به‌منظور دستیابی به دیدگاهی مشترک، توضیحاتی درخصوص برخی از آن‌ها ارایه می‌شود. یکی از پارامترهای مورد بررسی نوع استفاده از اینترنت است که در مورد هر یک توضیح داده می‌شود. دریافت اطلاعات، شامل دریافت اطلاعات درباره‌ی کالا یا خدمات، سلامت و خدمات پزشکی، سازمان‌های دولتی و غیر دولتی می‌شود. تعامل با سازمان‌های دولتی، شامل ارتباط دوطرفه با سازمان دولتی می‌شود. برای مثال می‌توان به دانلود فرم‌های ثبت نام کنکور سراری، کارت سوخت و غیره اشاره کرد. پست‌های الکترونیکی[۵]، شامل استفاده از پست‌های الکترونیک و دریافت یا ارایه‌ی اطلاعات از طریق آن است. تلفن زدن از طریق اینترنت، امکان استفاده از اینترنت به‌منظور مکالمه‌تلفنی را فراهم می‌نماید. ارسال اطلاعات یا پیام فوری، پیامی است که در محیط‌های گفتگو[۶] ارسال می‌شود. بانکداری اینترنتی، شامل پرداخت قبوض، اقساط وام‌های بانکی و غیره می‌شود. آموزش اینترنتی، شامل تحصیل در دانشگاه‌های مجازی، آموزش برخط[۷] زبان‌های خارجی، بسته‌های نرم‌افزاری و غیره می‌شود. خرید یا سفارش کالا، شامل خرید و فروش و سفارش کالا می‌شود. از دیگر موارد استفاده از اینترنت می‌توان به دانلود بازی کامپیوتری، فیلم، عکس، موسیقی، تماشای تلویزیون، گوش دادن به رادیو یا موسیقی، دانلود نرم‌افزار، خواندن یا دانلود روزنامه، مجله، یا کتاب به‌صورت برخط اشاره کرد. نحوه‌ی دسترسی به اینترنت، شامل Dialup، اینترنت هوشمند، ADSL و بی‌سیم می‌شود. همچنین در این مقاله، ویژگی‌های کاربران اینترنت بر اساس فعالیتی که انجام می‌دهند، مورد بررسی قرار گرفته است. این اطلاع از افراد ۱۰ ساله و بالاتر کسب شده است. شاغل به کسانی گفته

می‌شود که، در هفت روز گذشته حداقل یک ساعت کار کرده‌اند. کسانی که دارای شغلی هستند ولی در هفت روز گذشته به عللی از قبیل مرخصی و غیره کار نکرده‌اند و پس از رفع علت به کار خود ادامه خواهند داد نیز شاغل محسوب می‌شود. بیکار به کسانی که شاغل به‌حساب نمی‌آیند و در هفت روز گذشته در جستجوی کار بوده‌اند، اطلاق می‌شود. محصل به کسانی گفته می‌شود که شاغل و بیکار جویای کار محسوب نمی‌شوند و طبق برنامه‌های رسمی آموزشی کشور محصل تلقی می‌شوند. دارای درآمد بدون کار، شامل افراد شاغل، بیکار جویای کار و محصل نمی‌شود بلکه به کسانی گفته می‌شود که درآمدهای مستمری مانند حقوق بازنشستگی، حقوق وظیفه، درآمد املاک و مستغلات، سود سهام وغیره دارند. خانه‌دار، به کسانی گفته می‌شود که شاغل، بیکار جویای کار، محصل و دارای درآمد بدون کار به‌حساب نمی‌آیند و بنا بر اظهار پاسخگو، در هفت روز گذشته به خانه‌داری مشغول بوده‌اند. سایر برای افرادی است که از نظر وضع فعالیت در هیچ یک از گروه‌های بالا قرار نمی‌گیرند (افرادی که برای شرکت در کنکور آماده می‌شوند، سالمندان و غیره). در جدول ۱ متغیرهای مورد بررسی در این مقاله و نحوه‌ی دسته‌بندی و کدگذاری این متغیرها آورده شده است.

۴- استفاده از داده‌کاوی و رویکردهای آن در بررسی الگوهای رفتاری کاربران اینترنت

۴-۱- خوشه‌بندی کاربران اینترنت در سطح استان تهران

هدف از خوشه‌بندی، تفکیک مشاهدات به دسته‌های مجزا بر اساس شباهت داده‌ها به یکدیگر است؛ به‌طوری که اعضای هر دسته دارای بیش‌ترین شباهت به یکدیگر هستند و با اعضای دسته‌های دیگر متفاوت‌اند. تفاوت الگوریتم‌های مختلف این رویکرد بر اساس تعریف هر یک از آن‌ها از شباهت می‌باشد. K-means یکی از الگوریتم‌های شناخته‌شده‌ی این رویکرد است. الگوریتم‌های خوشه‌بندی برای متغیرهای کمی طراحی شده‌اند و در صورتی که متغیرهای مسئله، مانند متغیرهای این مطالعه از نوع کیفی[۸] باشند، نرم‌افزار کلمنتاین-۱۲ [۹] ابتدا این متغیرها را به متغیرهای عددی تبدیل کرده و سپس فاصله‌ی میان آن‌ها را محاسبه می‌کند [۲]. در این مطالعه، خوشه‌بندی داده‌ها از

جدول ۱- متغیرهای مورد بررسی

متغیرها	دسته‌بندی
جنسیت	(۱) مرد؛ (۲) زن.
سن	(۱) کم‌تر از ۱۸؛ (۲) ۳۰-۱۹؛ (۳) ۴۵-۳۱؛ (۴) ۶۰-۴۶؛ (۵) ۶۰ سال به بالا.
مدرک تحصیلی	(۱) ابتدایی، راهنمایی، متوسطه، دیپلم و پیش‌دانشگاهی، سوادآموز و غیر رسمی؛ (۲) کاردانی و کارشناسی، (۳) کارشناسی ارشد و دکتری حرفه‌ای؛ (۴) دکتری تخصصی.
وضع فعالیت	(۱) شاغل؛ (۲) بی‌کار؛ (۳) محصل؛ (۴) دارای درآمد بدون کار؛ (۵) خانه‌دار؛ (۶) سایر.
محل دسترسی به اینترنت	(۱) محل سکونت؛ (۲) محل تحصیل و کار؛ (۳) کافی‌نت، کتابخانه، منزل افراد دیگر و سایر، (۴) تلفن همراه و سایر دستگاه‌ها GPS، PDA.
تناوب دسترسی به اینترنت	(۱) حداقل یک‌بار در روز؛ (۲) حداقل یک‌بار در هفته ولی نه هر روز؛ (۳) حداقل یک‌بار در ماه ولی نه هر هفته؛ (۴) کم‌تر از یک‌بار در ماه.
نوع استفاده از اینترنت	(۱) دریافت اطلاعات در ارتباط با کالا و خدمات، سلامتی و خدمات پزشکی و سازمان‌های دولتی؛ (۲) آموزش اینترنتی؛ (۳) بازی و دانلود عکس، فیلم، نرم‌افزار، موسیقی و غیره و خواندن مجله و روزنامه و غیره؛ (۴) تعامل با سازمان‌های دولتی؛ (۵) فرستادن و دریافت ایمیل؛ (۶) تلفن زدن از اینترنت؛ (۷) ارسال اطلاعات یا پیام فوری؛ (۸) خرید یا سفارش کالا و خدمات؛ (۹) بانکداری اینترنتی.
نحوه‌ی دسترسی به اینترنت (برای افرادی که در محل سکونت از اینترنت استفاده می‌کنند)	(۱) Dialup و اینترنت هوشمند؛ (۲) ADSL؛ (۳) بی‌سیم وایمکس و انواع دیگر؛ (۴) سایر.

طریق نرم‌افزار کلمنتاین- ۱۲انجام شده است، داده‌ها به چهار خوشه تقسیم شدند، در این دسته‌بندی، تمایز میان خوشه‌ها جالب توجه بود و پس از ۱۳ بار تکرار خطای الگوریتم به ۰/۰ رسید. منظور از خطای الگوریتم مجموع مربعات خطا است که به‌صورت $\sum_{i=1}^{n}\sum_{j=1}^{k}\left\|x_i^{(j)}-c_j\right\|^2$ محاسبه می‌شود. در این رابطه فاصله میان پارامتر مورد

بررسی ($x_i^{(j)}$) از مرکز خوشه ($x_i^{(j)}$) محاسبه می‌شود. در جدول ۲ پارامترها به چهار خوشه تقسیم شده‌اند و درصد هر یک از پارامترها در این جدول نشان داده شده‌اند.

جدول ۲ نشان می‌دهد خوشه‌ی اول، شامل ۲۹۳ کاربر تهرانی است، ۵۷ درصد افراد این گروه زن هستند و تنها گروهی است که زن‌ها سهم بیش‌تری نسبت به مردها دارند. همچنین گروه سنی کم‌تر از ۳۱ تا ۴۵ سال درصد بیش‌تری را به خود اختصاص داده است و عمده‌ی افراد این گروه (۹۱ درصد) در گروه سنی ۱۹ تا کم‌تر از ۶۰ سال هستند. ۶۳ درصد افراد این خوشه، دارای سطح تحصیلات کاردانی و کارشناسی و ۳۴ درصد آن‌ها شاغل و ۲۸ درصد خانه‌دار هستند. بیش از ۷۰ درصد این افراد حداقل یک‌بار در ماه ولی نه هر هفته و عمدتاً (۹۷ درصد) در محل سکونت و با استفاده از Dial up یا اینترنت هوشمند (۹۲ درصد) از اینترنت استفاده کرده‌اند. در این گروه ۱۷ درصد افراد از دریافت اطلاعات بهره برده‌اند و ۹ درصد از افراد این خوشه برای بازی و دانلود عکس، فیلم و ... از اینترنت استفاده می‌کنند و بقیه‌ی موارد استفاده سهم بسیار کمی را به خود اختصاص داده است. جمعیت خوشه‌ی دوم ۲۹۶ نفر است. عمده‌ی افراد این گروه (۷۶ درصد) مرد هستند و بیش از نیمی از افراد این گروه در گروه سنی ۳۱ تا ۴۵ سال و بیش از ۸۰ درصد در گروه سنی ۱۹ تا ۴۵ سال قرار دارند. تقریباً تحصیلات نیمی (۵۳ درصد) از افراد این گروه کاردانی و کارشناسی بوده و ۹۰ درصد افراد این گروه شاغل هستند. ۹۰ درصد افراد این گروه در محل سکونت و محل کار و تحصیل از اینترنت استفاده می‌کنند. بیش از نیمی از افراد این گروه (۵۷ درصد) بیش از یک‌بار در روز و ۲۶ درصد حداقل یک‌بار در هفته ولی نه هر روز به اینترنت دسترسی دارند. نوع استفاده‌ی افراد این گروه به ترتیب بالاترین سهم استفاده، عبارت است از دریافت اطلاعات ۸۴ درصد، پست الکترونیک ۵۰ درصد، و تعامل با سازمان‌های دولتی ۴۲/۶ درصد و پس از آن بانکداری اینترنتی، بازی ودانلود قرار دارد. خوشه‌های سوم با ۳۹۱ نفر بیش‌ترین تعداد افراد را در خود جای داده است. ۶۰ درصد افراد این خوشه مرد هستند و عمده‌ی افراد این گروه (۹۴ درصد) در گروه سنی ۱۹ تا ۳۰ سال قرار دارند، ۶۵ درصد افراد این گروه دارای تحصیلات کاردانی و کارشناسی بوده و ۸۲ درصد افراد این گروه محصل هستند، این افراد عمدتاً در محل سکونت و تحصیل از اینترنت استفاده می‌کنند و بیش از نیمی از افراد این گروه حداقل یک‌بار در روز و ۲۲ درصد افراد حداقل یک‌بار در هفته و نه هر روز از اینترنت استفاده می‌کنند. نوع استفاده‌ی افراد این گروه عمدتاً سرگرمی (۷۷ درصد)، پست

الکترونیکی (۷۲ درصد) و دریافت اطلاعات (۶۱ درصد) است. ۵۰ درصد افراد از طریق Dial up و ۱۱ درصد از طریق ADSL در محل سکونت به اینترنت وصل می شوند. در خوشه‌های چهارم ۲۶۸ نفر قرار دارند که ۶۰ درصد آن‌ها مرد هستند و عمده‌ی افراد این گروه (۷۱ درصد) زیر ۱۸ سال سن دارند. با توجه به سن افراد گروه بدیهی است که ۹۵ درصد افراد دارای تحصیلات دیپلم و پائین‌تر و ۷۴ درصد آن‌ها محصل باشند، ۷۸ درصد افراد این گروه، در محل سکونت از اینترنت استفاده می‌کنند و عمدتاً به‌منظور سرگرمی (۷۹ درصد)، دریافت اطلاعات (۲۳ درصد)، پست الکترونیک (۱۷ درصد) و آموزش اینترنتی (۱۶ درصد) از اینترنت استفاده می‌کنند. در ضمن این افراد برای استفاده از اینترنت در محل سکونت از Dial up با ۶۹ درصد و ADSL با ۱۳ درصد استفاده می‌کنند.

جدول ۲- نتایج حاصل از خوشه‌بندی کاربران اینترنتی در سطح استان تهران

پارامترها	کل کاربران تهرانی	خوشه‌ی اول	خوشه‌ی دوم	خوشه‌ی سوم	خوشه‌ی چهارم
جنسیت					
مرد	۶۰/۱	۴۳/۳	۷۶/۰	۶۰/۰	۶۰/۸
زن	۳۹/۹	۵۶/۷	۲۴/۰	۴۰/۰	۳۹/۲
سن					
<۱۸	۱۶/۵	۱/۷	۰/۷	۲/۱	۷۱/۳
۱۹-۳۰	۴۸/۶	۲۹/۴	۳۳/۵	۹۳/۹	۲۰/۲
۳۱-۴۵	۲۲/۴	۳۳/۱	۵۲/۴	۳۳/۱	۶/۰
۴۶-۶۰	۱۰/۳	۲۸/۷	۱۲/۵	۰/۸	۱/۹
۶۰>	۲/۱	۷/۲	۱/۰	۰/۸	۰/۸
مدرک تحصیلی					
دیپلم و پایین‌تر	۳۶/۰	۳۱/۷	۲۸/۰	۴/۹	۹۴/۸
کاردانی و کارشناسی	۴۹/۰	۶۳/۴	۵۳/۷	۶۵/۲	۴/۵
کارشناسی ارشد و دکتری حرفه‌ای	۱۳/۱	۴/۸	۱۵/۲	۲۶/۳	۰/۴
دکتری تخصصی	۲/۰	۰/۳	۳/۰	۳/۶	۰/۴

ادامه‌ی جدول ۲- نتایج حاصل از خوشه‌بندی کاربران اینترنتی در سطح استان تهران

پارامترها	کل کاربران تهرانی	خوشه‌ی اول	خوشه‌ی دوم	خوشه‌ی سوم	خوشه‌ی چهارم
وضع فعالیت					
شاغل	۳۵/۶	۳۳/۸	۹۰/۹	۱۰/۷	۱۲/۷
بی‌کار	۴/۵	۸/۵	۲/۴	۳/۶	۳/۷
محصل	۴۴/۰	۷/۹	۳/۰	۸۱/۶	۷۳/۵
دارای درآمد بدون کار	۰/۵	۱۸/۱	۱/۴	۰/۳	۱/۵
خانه‌دار	۸/۳	۲۸/۳	۱/۴	۱/۸	۳/۴
سایر	۲/۴	۳/۴	۱/۰	۲/۱	۳/۴
محل دسترسی به اینترنت					
محل سکونت	۶۱/۵	۹۶/۹	۳۵/۵	۴۳/۵	۷۷/۶
محل کار و تحصیل	۷/۷	۱/۰	۲۱/۰	۶/۱	۲/۶
کافی نت، کتابخانه و منزل افراد دیگر	۴/۱	۰/۳	۲/۷	۳/۳	۱۰/۸
تلفن همراه و سایر دستگاه‌ها	۱/۰	۰/۰	۱/۴	۱/۰	۱/۵
محل سکونت و کار و تحصیل	۱۳/۹	۰/۷	۳۳/۸	۱۷/۴	۱/۱
محل سکونت و کافی‌نت و کتابخانه و ...	۲/۶	۱/۰	۰/۳	۴/۴	۴/۱
محل سکونت و تلفن همراه	۰/۲	۰/۰	۰/۳	۰/۵	۰/۰
محل کار و تحصیل و کافی‌نت و ...	۱/۳	۰/۰	۰/۷	۳/۱	۰/۸
محل کار و تحصیل و تلفن همراه و ...	۰/۲	۰/۰	۰/۰	۰/۵	۰/۰
کافی‌نت، کتابخانه و تلفن همراه و ...	۰/۲	۰/۰	۰/۰	۰/۳	۰/۸
محل سکونت، کار و تحصیل و کافی‌نت و ...	۵/۵	۰/۰	۲/۴	۱۵/۶	۰/۴

ادامه‌ی جدول ۲- نتایج حاصل از خوشه‌بندی کاربران اینترنتی در سطح استان تهران

پارامترها	کل کاربران تهرانی	خوشه‌ی اول	خوشه‌ی دوم	خوشه‌ی سوم	خوشه‌ی چهارم
محل سکونت، کار و تحصیل و تلفن همراه	۰/۵	۰/۰	۰/۳	۱/۳	۰/۰
محل سکونت، کافی‌نت و کتابخانه و تلفن همراه و ...	۰/۳	۰/۰	۰/۳	۰/۵	۰/۴
محل سکونت، کار و تحصیل، کافی‌نت و کتابخانه و تلفن همراه و ...	۱/۰	۰/۰	۰/۷	۲/۶	۰/۰
تناوب دسترسی به اینترنت					
حداقل یک‌بار در روز	۳۶/۷	۴/۴	۵۷/۸	۵۴/۲	۲۳/۱
حداقل یک‌بار در هفته ولی نه هر روز	۲۲/۴	۱۱/۶	۲۶/۰	۲۲/۰	۳۱/۰
حداقل یک‌بار در ماه ولی نه هر هفته	۲۸/۵	۷۷/۵	۷/۴	۹/۰	۲۶/۵
کمتر از یک‌بار در ماه	۱۲/۴	۶/۵	۸/۸	۱۴/۸	۱۹/۴
نوع استفاده از اینترنت					
دریافت اطلاعات	۴۸/۲	۱۷/۱	۸۴/۸	۴۰/۹	۲۳/۱
ارسال اطلاعات یا پیام	۱۳/۱	۳/۴	۱۴/۲	۲۴/۰	۶/۷
پست الکترونیک	۳۹/۰	۳/۴	۵۰/۰	۷۲/۴	۱۶/۸
تعامل با سازمان‌های دولتی	۱۹/۵	۴/۸	۴۲/۶	۲۰/۰	۹/۳
بانک‌داری اینترنتی	۱۴/۹	۳/۴	۲۸/۰	۲۰/۷	۴/۵
آموزش اینترنتی	۱۲/۶	۳/۸	۷/۸	۲۰/۸	۱۵/۷
خرید یا سفارش کالا و خدمات	۸/۸	۲/۱	۱۵/۵	۱۲/۳	۳/۷

ادامه‌ی جدول ۲- نتایج حاصل از خوشه‌بندی کاربران اینترنتی در سطح استان تهران

پارامترها	کل کاربران تهرانی	خوشه‌ی اول	خوشه‌ی دوم	خوشه‌ی سوم	خوشه‌ی چهارم
سرگرمی	۴۹/۰	۹/۲	۲۳/۳	۷۷/۸	۷۹/۱
تلفن اینترنتی	۲/۶	۱/۴	۵/۷	۲/۶	۰/۸
نحوه‌ی دسترسی به اینترنت					
Dial up و اینترنت هوشمند	۶۷/۱	۹۲/۲	۵۹/۱	۵۰/۴	۶۹/۴
ADSL	۱۰/۳	۳/۸	۱۳/۹	۱۰/۵	۱۳/۱
بیسیم وایمکس و انواع دیگر	۰/۹	۰/۰	۲/۴	۰/۳	۱/۱
Dial up و اینترنت هوشمند وADLS	۰/۱	۰/۳	۰/۰	۰/۰	۰/۴
سایر	۱/۴	۰/۰	۲/۰	۲/۳	۰/۸

۲-۴- قوانین انجمنی

این قوانین که یکی از پرکاربردترین رویکردهای روش‌های یادگیری بدون ناظر است و به یافتن روابطی میان ویژگی‌های داده‌هامی‌پردازد و به دنبال تحلیل وابستگی‌ها، مطالعه‌ی ویژگی‌ها و خصوصیاتی است که با یکدیگر همراه هستند. الگوریتم‌های موجود به دنبال استخراج قوانین به‌منظور کاهش ارتباط میان دو یا چند خصوصیت هستند. بیش‌ترین کاربرد قوانین انجمنی در اقتصاد و به‌خصوص تحلیل سبد خرید[۱۰] است. این قوانین که به‌صورت «اگر[۱۱]» و «آن‌گاه[۱۲]» بیان می‌شوند، می‌توانند الگوها و رفتارهای جالبی را میان برخی از متغیرهای موجود در مجموعه‌ی داده‌ها نشان دهند. همچنین برای نشان دادن میزان سودمندی و اطمینان روابط بدست آمده دو معیار پشتیبان[۱۳] و اطمینان[۱۴] نیز تعریف می‌شوند [۳]. اگر قانون استخراج‌شده به‌صورت $X \to Y$ باشد، معیار پشتیبان نشان‌دهنده‌ی درصد یا تعداد مجموعه‌ی تراکنش‌هایی در کل مجموعه است که شامل هر دو مجموعه‌ی X و Y باشد و معیار اطمینان نیز میزان وابستگی یک مشخصه‌ی خاص را به مشخصه‌ی دیگر بیان می‌نماید.

$$confidence\ (Y|X) = \sup port(X \cup Y)/\sup port(X).$$

یکی از شناخته‌شده‌ترین الگوریتم‌های قوانین انجمنی، الگوریتم اپریوری[15] است. در این مطالعه برای یافتن قوانین میان پارامترهای مسئله از این الگوریتم در نرم افزار کلمنتاین- ۱۲ بهره گرفته شده است [۲].

برای اجرای این الگوریتم در نرم‌افزار، ابتدا تمامی متغیرها هم به‌صورت متغیر ورودی و هم خروجی در نظر گرفته شدند. حداقل معیار پشتیبان برابر با ۱۰ درصد و حداقل میزان اطمینان برابر با ۷۰ درصد قرار داده شد. با اجرای الگوریتم اپریوری ، تعداد ۳۹/۱۶۹ قانون بدست آمد، که پس از بررسی‌های انجام‌شده، تعدادی از این قوانین که بیانگر ویژگی‌های جالب توجه کاربران اینترنت در استان تهران بودند، استخراج گردیدند. جدول ۳ بیانگر تعدادی از این قوانین می‌باشد.

۵- تجزیه و تحلیل نتایج

در این بخش ابتدا به بررسی ویژگی‌های جمعیتی کاربران اینترنت در سطح استان تهران پرداخته شده، سپس نتایج بدست آمده از خوشه‌بندی و ویژگی‌های هر خوشه مورد مطالعه قرار گرفته است.

بررسی ویژگی‌های کاربران اینترنت نشان می‌دهد، ۶۰ درصد کاربران مرد هستند و عمدتاً در گروه‌های سنی جوان و میانسال (۱۹ تا ۴۵ سال) قرار دارند، ۸۰ درصد این افراد محصل و شاغل بوده و تقریباً نیمی از آن‌ها دارای مدرک کاردانی و کارشناسی هستند. کاربران تهرانی عمدتاً در محل سکونت خود به اینترنت دسترسی دارند. ۷۷ درصد آن‌ها در محل سکونت با Dial up، اینترنت هوشمند و ADSL به اینترنت متصل می‌شوند که Dial up و اینترنت هوشمند سهم عمده‌ای دارد و کاربران تهرانی عمدتاً به‌منظور دریافت اطلاعات (۴۸ درصد)، بازی و دانلود (۴۹ درصد) و پست الکترونیک (۳۹ درصد) از اینترنت استفاده می‌کنند.

با استفاده از نتایج جدول ۲ به ارایه‌ی عمده‌ی ویژگی‌های هر خوشه، پرداخته می‌شود. خوشه‌ی اول را عمدتاً زنانی بین ۳۱ تا ۴۵ سال با تحصیلات کاردانی و کارشناسی تشکیل می‌دهد که حداقل یک‌بار در ماه و نه هر هفته به‌منظور دریافت اطلاعات و دانلود

جدول ۳- نتایج بدست آمده از قوانین انجمنی برای کاربران اینترنت در سطح استان تهران

اطمینان	پشتیبان	اگر	آن‌گاه
۹۹/۴	۱۰/۴	سن: زیر ۱۸ سال خدمات دولتی: عدم استفاده وضع فعالیت: محصل	کسب و کار اینترنتی: عدم استفاده
۹۱/۵	۱۰/۹	تناوب دسترسی: حداقل یک‌بار در روز محل دسترسی: محل سکونت و سایر نقاط جنسیت: مرد بانکداری: عدم استفاده	وضع فعالیت: محصل
۹۱/۵	۱۰/۲	نحوه‌ی دسترسی: Dial up و هوشمند جنسیت: مرد دریافت اطلاعات: استفاده	خرید و یا سفارش کالا: عدم استفاده
۹۱/۲	۱۰/۱	مدرک تحصیلی: دیپلم و پایین‌تر نحوه‌ی دسترسی: Dial up و هوشمند پست الکترونیک: عدم استفاده	آموزش اینترنتی: عدم استفاده
۷۰/۱	۱۳/۴	تناوب دسترسی: حداقل یک‌بار در روز دریافت اطلاعات: استفاده آموزش اینترنتی: عدم استفاده	پست الکترونیک: استفاده
۷۰/۳	۱۷/۳	محل دسترسی: محل سکونت تعامل دولتی: عدم استفاده آموزش اینترنتی: عدم استفاده بانک‌داری اینترنتی: عدم استفاده	نحوه‌ی دسترسی: Dial up
۷۰/۳	۱۱/۳	مدرک تحصیلی: دیپلم و پایین‌تر دریافت اطلاعات: استفاده آموزش اینترنتی: عدم استفاده بانک‌داری اینترنتی: عدم استفاده	نحوه‌ی دسترسی: Dial up

عکس و فیلم از اینترنت استفاده می‌کنند. خوشه‌ی دوم را عمدتاً مردانی بین ۳۱ تا ۴۵ سال با تحصیلات کاردانی و کارشناسی تشکیل می‌دهد که شاغل بوده و حداقل یک‌بار در روز یا حداقل یک‌بار در هفته و نه هر روز در محل سکونت، کار و تحصیل به‌منظور دریافت اطلاعات، پست الکترونیک و تعامل با سازمان‌های دولتی از اینترنت استفاده می‌کنند. خوشه‌ی سوم را عمدتاً مردانی بین ۱۹ تا ۳۱ سال با تحصیلات کاردانی و

کارشناسی تشکیل می‌دهد که محصل بوده و حداقل یک‌بار در روز یا حداقل یک‌بار در هفته و نه هر روز به‌منظور دریافت اطلاعات، پست الکترونیک، بازی و دانلود عکس و فیلم از اینترنت استفاده می‌کنند. خوشه‌ی چهارم عمدتاً مردان زیر ۱۸ سال با تحصیلات دیپلم و پایین‌تر هستند که محصل بوده و به‌منظور دریافت اطلاعات، پست الکترونیک و آموزش اینترنتی، بازی و دانلود عکس و فیلم از اینترنت استفاده می‌کنند.

کاربران اینترنت در استان تهران، در مجموع و به‌تفکیک خوشه‌ها از خرید و سفارش کالا و تلفن زدن اینترنتی کم‌ترین استفاده را داشته‌اند که نشان از عدم گسترش خرید اینترنتی در کشور است، که این امر از طرف عرضه کنندگان محصولات و فروشگاه‌ها نیز مورد توجه نیست و تبلیغاتی روی آن انجام نمی‌شود، در حالی که گسترش این امر و توجه به بانک‌داری و آموزش اینترنتی در شهری مانند تهران، حجم بالایی از رفت و آمدهای داخل شهری را کاهش می‌دهد.

قوانین انجمنی مندرج در جدول ۳ نشان می‌دهد، استفاده از Dial up مانعی برای بهره‌گیری کاربران از خدماتی همچون بانک‌داری و آموزش اینترنتی است. همچنین سطح تحصیلات افراد نیز می‌تواند در این مسئله تأثیرگذار باشد.

۶- بحث و نتیجه‌گیری

با توجه به نفوذ و گسترش اینترنت در سطح کشور، بررسی نحوه‌ی استفاده و ویژگی‌های کاربران اینترنت یکی از مسائل جالب توجه می‌باشد، که تاکنون مورد بررسی قرار نگرفته است. این مطالعه، به بررسی ویژگی‌های کاربران تهرانی - بر اساس اطلاعات آخرین طرح آمارگیری از کاربران اینترنت- پرداخته است. برای یافتن الگوهای رفتاری و ویژگی‌های کاربران از دو رویکرد کاربردی داده‌کاوی یعنی خوشه‌بندی و قوانین انجمنی استفاده شده است. نتایج این تحقیق نشان می‌دهد، به‌ترتیب بازی و دانلود، دریافت اطلاعات و پست الکترونیک بیش‌ترین سهم استفاده از اینترنت را به خود اختصاص داده است. پس از آن تعامل با سازمان‌های دولتی و بانکداری اینترنتی قرار دارد که گسترش این امر به‌دلیل تلاش‌های دولت برای گسترش دولت الکترونیک و عدم ارایه‌ی حضوری برخی خدمات همچون ثبت‌نام آزمون‌ها، کارت سوخت، یارانه و غیره است. اما گسترش خدماتی نظیر بانک‌داری و آموزش اینترنتی نیازمند تلاش‌های بیش‌تری از جانب مسئولین، مؤسسات مالی و آموزشی و ارایه‌دهندگان سرویس‌های اینترنتی می‌باشد. ارتقای آگاهی شهروندان

از میزان امنیت شبکه‌ی بانکی، ساده‌سازی و آموزش فرایند ثبت تراکنش‌ها می‌تواند به افزایش استفاده از بانک‌داری اینترنتی در سطح جامعه منجر شود. در ضمن گسترش فرهنگ خرید اینترنتی میان هر دو طرف عرضه‌کنندگان و تقاضاکنندگان و ایجاد بسترهای مناسب می‌تواند به کاهش رفت و آمدها در تهران کمک نماید.

توضیحات

1. Supervised learning
2. Unsupervised learning
3. Clustering
4. Association rules
5. E-mail
6. Chat
7. On-Line
8. Categorical
9. Clementine 12
10. Market basket analysis
11. If
12. Then
13. Support
14. Confidence
15. Apriori

مرجع‌ها

[۱] مرکز آمار ایران (۱۳۸۹).نشریه‌ی نتایج طرح آمارگیری از کاربران اینترنت، تهران.

[2] Clementine® 12.0 Algorithms Guide; *http://www.spss.com*

[3] Han, J. and Kamber, M. (2006). *Data Mining Concept and Techniques*, second edition.Morgan Kaufmann Publishers is an imprint of Elsevier; San Francisco, CA.

[4] Okazaki, S. (2006).What do we know about mobile internet adopters? A cluster analysis. *Information and Management*, **43**, 127-141.

[5] Pietquin, O. (2004).*A Framework for Unsupervised Learning of Dialogue Strategies*; Presses universitaires de Louvain; Belguim.

سحر صاحبی
کارشناس ارشد علوم اقتصادی
تهران، خیابان فاطمی، نبش خیابان رهی معیری، مرکز آمار ایران.
رایانشانی: saharsahebi@gmail.com

سپیده صاحبی
کارشناس ارشد مهندسی سیستم‌های اقتصادی و اجتماعی
تهران، بلوار آفریقا، خیابان شهید عاطفی شرقی، شرکت ملی نفتکش ایران.
رایانشانی: sahebi@ptsoc.com