پژوهش‌های مدیریت منابع سازمانی
دوره ۶، شماره ۴، زمستان ۱۳۹۵

ارائه الگویی برای تحلیل رفتار کاربران شبکه‌های اجتماعی با استفاده از روش‌های داده‌کاوی: یک شبکه اجتماعی در ایران

بابک سهرابی[۱]، ایمان رئیسی وانانی[۲]، مرضیه طالبیان[۳]

۱- استاد گروه مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه تهران، تهران، ایران
۲- استادیار گروه مدیریت صنعتی، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبایی، تهران، ایران
۳- دانشجوی کارشناسی ارشد مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه تهران، تهران، ایران

دریافت: ۱۳۹۵/۱/۱۵ پذیرش: ۱۳۹۵/۵/۱۶

چکیده

در فضای مجازی، شبکه‌های اجتماعی به عنوان نوع جدیدی از وب‌سایت‌ها متولد شده‌اند و کاربردها، کاربران و طرفداران بسیاری پیدا کرده‌اند. شبکه‌های اجتماعی یکی از انوع رسانه‌های اجتماعی محبوب محسوب می‌شوند و محلی برای شکل‌گیری جماعت‌های مجازی و شبکه‌سازی کاربران اینترنتی هستند کاربران اینترنتی را برحسب نوع مواجه خود با شبکه‌های اجتماعی می‌توان به شکل‌های مختلف دسته‌بندی کرد. باتوجه به گسترش انواع شبکه‌های اجتماعی، نیاز به الگویی است تا براساس آن تصمیم‌گیری استراتژیک و یا اتخاذ خط مشی‌های جدید برای خدمت‌رسانی بهتر به کاربران صورت گیرد. هدف این پژوهش، ارائه سازوکاری جهت پیش‌بینی الگوهای رفتاری افراد در شبکه‌های اجتماعی و به‌کارگیری تکنیک‌های داده‌کاوی با استفاده از روش فرایندی اجرای پروژه‌های داده‌کاوی برای رده‌بندی و تحلیل رفتار کاربران شبکه اجتماعی به منظور شناخت بهتر آنها و در نتیجه بهبود خدمات ارائه شده و تدوین استراتژی‌های مناسب می‌باشد. جامعه آماری پژوهش کاربرانی می‌باشد که از شبکه اجتماعی مورد نظر استفاده می‌کنند که شامل تعداد ۳۱۰۳۳ کاربر فعال است. درک صحیح از الگوهای

* نویسنده مسئول مقاله E-mail: bsohrabi@ut.ac.ir

رفتاری کاربران شبکه‌های اجتماعی، منجر به انطباق هرچه بهتر خدمات ارائه شده به‌وسیله شبکه با نیازهای کاربر و به تبع آن، توسعه تعداد کاربران شبکه و افزایش ارزش افزوده آن برای کاربران و درآمدزایی برای متولیان شبکه می‌گردد.

واژه‌های کلیدی: شبکه اجتماعی، داده‌کاوی، رفتار کاربران، خوشه‌بندی، دیویس بولدین.

۱- مقدمه

در فضای مجازی شبکه‌های اجتماعی به عنوان نوع جدیدی از وب‌سایت‌ها متولد شده‌اند و کاربران و طرفداران زیادی پیدا کرده‌اند. شبکه‌های اجتماعی یکی از انوع رسانه‌های اجتماعی محسوب می‌شوند و محلی برای شکل‌گیری جماعت‌های مجازی و شبکه‌سازی کاربران اینترنتی هستند. شبکه‌های اجتماعی موفق شده‌اند تعداد قابل توجهی از کاربران اینترنتی را جذب کنند و در سال گذشته میلادی در میان چهار فعالیت اصلی کاربران اینترنتی قرار گرفتند. برخی تحلیلگران وب، آینده فضای مجازی را در اختیار این نوع سایت‌ها می‌دانند.

صدها سایت شبکه اجتماعی با حوزه‌های تخصصی متفاوت و همچنین با زبان‌های متنوع برای کاربران کشورهای مختلف در اینترنت فعالیت می‌کنند.کاربران شبکه های اجتماعی می‌توانند در این سایت‌ها صفحات و پروفایل‌های شخصی برای خودشان ایجاد کنند، شبکه‌ای مجازی از دوستان خود پدید آورند؛ آنها می‌توانند همانند فضایی که وبلاگ‌ها و میکروبلاگ‌ها در اختیارشان قرار می‌دهند یادداشت‌های کوتاه و بلندشان را منتشر کنند؛ عکس، صدا و ویدیوهای شخصی خود را آپلود کنند؛ از آخرین اخبار و رویدادها در حوزه‌های مختلف آگاه شوند؛ در صفحات هواداری و اتاق‌های گفتگوی متنوع عضو شوند و قابلیت‌های فراوان دیگری که ممکن است هر شبکه اجتماعی برای کاربرانش ایجاد کند. کاربران شبکه‌های اجتماعی از این قابلیت‌ها و امکانات متنوع به یک میزان و در یک سطح استفاده نمی‌کنند. برخی کاربران اینترنتی در چند شبکه اجتماعی عضو هستند، روزانه به این سایت‌ها سر می‌زنند و اکثر امکانات آنها را به کار می‌گیرند و برخی دیگر ممکن است تنها در یک سایت عضو باشند و هر از چندگاهی تنها به صفحه شخصی خودشان مراجعه کنند. بین این دو گروه نیز کاربران اینترنتی برحسب میزان و نوع استفاده خود از شبکه‌های اجتماعی در طیف وسیعی

قرار می‌گیرند. کاربران اینترنتی را برحسب نوع مواجه‌شان با شبکه‌های اجتماعی می‌توان به شکل‌های مختلف دسته‌بندی کرد.

داده‌کاوی را می‌توان شناخت رابطه‌های منطقی و الگوهای موجود و روابط پنهان میان داده‌ها دانست که در جوامع مختلف به دنبال یافتن الگوهای مفید می‌باشد. در این پژوهش هدف این است که به تحلیل رفتار کاربران شبکه‌های اجتماعی پرداخته شود. منظور از تحلیل رفتار کاربران این می‌باشد که پس از پیش‌پردازش داده‌ها با استفاده از تکنیک‌های داده‌کاوی مدل داده‌ای طراحی شود تا بتوان بین کاربران شبکه‌های اجتماعی تفکیک قائل شد و آنها را در خوشه‌های مختلف تقسیم‌بندی کرد تا در آینده براساس دانش حاصل از این مدل در اتخاذ تصمیم مناسب در مقابل رفتار کاربران مورد استفاده قرار گیرد.

۲- پیشینه تحقیق

۲-۱- پیشینهٔ نظری

۲-۱-۱- شبکه اجتماعی

در علوم اجتماعی، میل جمعی به شرکت در یک جامعه پدیده‌ای است که مدت‌ها مورد بررسی قرار گرفته‌است [۱]. حدود چهارصد سال قبل ارسطو انسان را به عنوان شخصیتی که نیاز اساسی به جستجو و ایجاد جوامع دارد، توصیف کرده است [۲]. بنابراین ایده کلی شبکه‌های اجتماعی جدید نیست. با این حال با ظهور شبکه جهانی وب و توسعه فناوری اطلاعات، شبکه‌های اجتماعی به یک بعد جدید دست پیدا کردند. به لطف انواع مختلف نرم‌افزارهای اجتماعی مانند وبلاگ‌ها، سایت‌های تولید محتوا توسط کاربر و جوامع مجازی در سراسر شبکه جهانی وب، مردم شروع به اتصال و برقراری ارتباط آنلاین با یکدیگر نمودند [۲]. همراه با این تغییرات کاربرانی که پیش از این غیرفعال بوده‌اند، به تولیدکنندگان محتوا در شبکه جهانی وب تبدیل شدند[۳]. در کنار این تغییرات، وب ۲,۰ نیز شناخته شد. شبکه‌های اجتماعی نیز به عنوان یک رسانه جمعی رایگان و جدید به کاربران به شکل گسترده‌ای ارائه شد.

۲-۱-۲- تعریف شبکه‌های اجتماعی

شبکه‌های اجتماعی نوع خاصی از جوامع مجازی است [۴]. با این حال، با توجه به اینکه یک پدیده جدید مرتبط با وب ۲,۰ می‌باشد، در نتیجه نه یک اصطلاح پذیرفته شده و نه تعریف تثبیت شده‌ای از ان وجود ندارد. بوید و الیسون شبکه‌های اجتماعی را به این صورت تعریف کرده‌اند که «سایت‌های شبکه اجتماعی به عنوان خدمات مبتنی بر وب است که به افراد اجازه می‌دهد (۱)، یک پروفایل عمومی یا نیمه عمومی در یک سیستم محدود را بسازد؛ (۲)، فهرستی از سایر کاربران برای اشتراک‌گذاری ارائه دهد؛ (۳)، فهرست افراد مرتبط با وی و کسانی را که از طریق سایر افراد سیستم به او معرفی شده‌اند، ببیند و اوقاتش را سپری سازد» [۵].

این اصطلاحات متفاوت برای شبکه‌های اجتماعی آنلاین اغلب به صورت مترادف یکدیگر مورد استفاده قرار می‌گیرند، حتی اگر یک تعریف مشترکی از موضوع مورد نظر را به اشتراک نگذارند، برای مثال بوید و الیسون اشاره می‌کنند که آنها به طور عمد اصطلاح «شبکه‌سازی» را جهت تأکید بر شروع رابطه بین افراد غریبه انتخاب نکرده‌اند. درحالی که چنین شبکه‌سازی در این سایت‌ها امکان‌پذیر می‌باشد، وظیفه اصلی آنها نیست. نمونه چنین سایت‌های محتواگرایی هم یوتیوب و توئیتر می‌باشد. بیر [۶] معتقد است تعریفی که بوید و الیسون ارائه داده‌اند، بسیار گسترده است. بنابرین شبکه‌های اجتماعی آنلاین را برطبق نظر بوید و الیسون تعریف کرده اما تمرکز بر سایت‌های کاربر محور می‌باشد.

۲-۱-۳- داده‌کاوی

رشد و نفوذ کامپیوتر در سیستم‌های اجتماعی و اقتصادی، قابلیت آنها را در تولید و نگهداری داده از منابع مختلف ارتقا داده است. در چنین شرایطی، حجم بسیار بالایی از داده‌ها در مورد تمامی جنبه‌های سیستم‌ها تولید شده است. این رشد سریع حجم داده‌ها، نیاز مبرمی به تکنیک‌ها و ابزارهای اتوماتیک برای تبدیل داده‌ها به اطلاعات و دانش را ایجاد کرده است. این قضیه منجر به وجود آمدن حوزه جدیدی در علوم کامپیوتر به نام داده‌کاوی شده است. تعریف‌های متفاوتی از داده‌کاوی وجود دارد ولی تعریفی که در بیشتر مراجع به اشتراک ذکر

شده عبارت است از "استخراج اطلاعات و دانش و کشف الگوهای پنهان از یک پایگاه داده بسیار بزرگ و پیچیده" [۷، ص ۲۷۴].

داده‌کاوی کمک می‌کند تا سازمان‌ها با کاوش بر داده‌های یک سیستم، الگوها، روندها و رفتارهای آینده را کشف و پیش‌بینی کرده و بهتر تصمیم بگیرند. داده‌کاوی با استفاده از تحلیل وقایع گذشته یک تحلیل خودکار و پیش‌بینانه ارائه می‌کند و به سؤالاتی جواب می‌دهد که پاسخ آنها در گذشته ممکن نبوده و یا به زمان زیادی نیاز داشت. ابزارهای داده‌کاوی الگوهای پنهانی را کشف و پیش بینی می‌کنند که متخصصان ممکن است به دلیل اینکه این اطلاعات و الگوها خارج از انتظار آنها باشد، آنها را مد نظر قرار ندهند و به آنها دست پیدا نکنند. الگوهای استخراج شده می‌توانند رابطه‌ای بین ویژگی‌ها و مشخصات سیستم مانند نوع تقاضا و نوع مشتری، پیش‌بینی‌های آینده براساس مشخصات سیستم، قوانین (اگر – آن گاه) بین متغیرهای سیستم، دسته‌بندی‌ها و خوشه‌بندی‌های اشیا و رکوردهای شبیه به هم در یک سیستم و غیره باشند [۸].

پیش‌پردازش زمان برترین مرحله فرایند کشف دانش است ولی با توجه به تأثیر مستقیم آماده‌سازی داده‌ها بر کیفیت نتایج داده‌کاوی، اجرای این مرحله ضروری می‌باشد. داده‌های موجود در دنیای واقعی ممکن است کیفیت لازم برای شروع داده‌کاوی را نداشته باشند، برای مثال وجود« نویز[1]»، «نمونه‌های پرت[2]»، «مقادیر از دست رفته[3]»و داده‌های تکراری در داده‌ها، اجرای مرحله پیش‌پردازش را ضروری می کند. همچنین ممکن است به علت جمع‌آوری داده از پایگاه داده‌های مختلف این داده ها دارای فرمت‌های متفاوتی باشند. انجام داده‌کاوی روی داده‌هایی که دارای کیفیت پایین هستند، منجر به دستیابی به نتایج با کیفیت پایین خواهد شد.

بنابراین می‌توان گفت اجرای مرحله پیش‌پرازش روی داده‌ها قبل از داده‌کاوی عملکرد کل فرایند را بهبود بخشد. از این رو باید به انتخاب روش‌های مناسب برای پیش‌پردازش توجه خاصی شود. روش مورد استفاده در این پژوهش برای پیش پردازش داده‌ها از شاخص «عامل پرت محلی [4]»استفاده شده است . این روش یکی از محبوب‌ترین رویکردهای تشخیص

1. Noise
2. Outlier
3. Missing Value
4. Local Outlier Factor

داده‌های پرت مبتنی بر چگالی است [۹]. نمره این الگوریتم براساس نسبت تراکم قابل دسترسی محلی از k همسایگی از شیء o بررسی می‌شود. این تراکم قابل دسترسی که برای محاسبه عامل پرت محلی استفاده می‌شود، فاکتوری هم برای k نزدیک‌ترین همسایگی شیء o و اندازه فاصله قابل دسترسی می‌باشد [۱۰].

از نظر مفهومی خوشه‌بندی، یعنی گروه بندی یک سری موجودیت در گروهای مختلف به طوری که این گروه‌ها نشان‌دهنده مفهوم یا معنی خاصی باشند و یا به عبارت ساده‌تر به یکدیگر شبیه باشند. خوشه‌بندی روشی آماری است که به مقایسه کمّی تعدادی موجودیت براساس ویژگی‌های آنها پرداخته و گروه‌های مختلفی را که آن موجودیت‌ها به آن تعلق دارند، اکتشاف می‌کند. به بیان دیگر، خوشه‌بندی، یعنی دسته‌بندی داده‌ها به *k* گروه مختلف به طوری که داده‌هایی که در یک دسته قرار می‌گیرند به یکدیگر شبیه باشند و داده‌های دسته‌های مختلف با یکدیگر تفاوت داشته باشند [۱۱].

۲-۱-۴- انواع روش‌های خوشه‌بندی

روش‌های خوشه‌بندی در داده‌کاوی عبارتند از :

- روش‌های سلسله مراتبی مانند: Average-Linkage , Single-Linkage
- روش‌های برپایه مرکز خوشه مانند : Fuzzy c-means , K-medoids ,k-means
- روش‌های بر پایه توزیع مانند : Excepion-Maximization (EM)
- روش‌های بر پایه چگالی مانند : Dbscan , Optics

در این پژوهش از روش k میانگین جهت خوشه‌بندی داده‌ها استفاده شده است. الگوریتم خوشه‌بندی k میانگین از جمله مشهورترین الگوریتم‌های یادگیری بدون نظارت است که در آن مجموعه داده‌ها به تعداد خوشه‌های از پیش تعیین شده تقسیم می‌شوند. این الگوریتم از روش خوشه‌بندی افرازی استفاده می‌کند. مراحل مختلف الگوریتم خوشه‌بندی k میانگین به صورت زیر می‌باشد:

۱. نخست K نقطه به صورت تصادفی به عنوان مراکز خوشه انتخاب می‌شوند.

۲. هر رکورد در مجموعه داده به خوشه‌ای که مرکز آن خوشه کمترین فاصله تا آن رکورد را دارا است ، نسبت داده می‌شود. مشهورترین معیارهای محاسبه فاصله رکوردها در روش‌های

خوشه‌بندی معیارهای فاصله اقلیدسی و فاصله همینگ هستند که به‌ترتیب در معادلات فرمول ۱ و فرمول ۲ ارائه شده‌اند. در این معادلات n بیانگر تعداد مشخصه‌ها و x_k و y_k به‌ترتیب k امین ویژگی رکورد x و y هستند.

فرمول ۱

$$d_E(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

فرمول ۲

$$d_H(x,y) = \sum_{k=1}^{n}|x_k - y_k|$$

۳. پس از تخصیص تمام رکوردها به یکی از خوشه‌های تشکیل شده، برای هر خوشه یک نقطه جدید به عنوان مراکز خوشه محاسبه می‌شود.

۲-۱-۵- ارزیابی خوشه‌بندی

پس از ساخت مدل، هر تحلیلگر داده‌کاوی باید نتایج را به صورت شهوردی بررسی کرده و با استفاده از دامنه دانش خود، نتایج را تفسیر کند. در این مرحله می‌توان از کمک صاحبنظران حوزه پروژه استفاده کرد و به‌وسیله آنها نتایج حاصل از مدلسازی را بررسی کرده و این نتایج را با شواهد و قرائن و همچنین دانش قبلی مقایسه نمود و از صحت شهودی مدل آگاه شد. برای روش‌های تشریحی داده‌کاوی نظیر خوشه‌بندی، نتایج حاصل از اعمال الگوریتم‌های خوشه‌بندی روی یک مجموعه داده با توجه به انتخاب‌های پارامترهای الگوریتم‌ها می‌تواند بسیار متفاوت از یکدیگر باشد. هدف از اعتبارسنجی خوشه‌ها یافتن خوشه‌هایی است که بهترین تناسب با داده‌های مورد را داشته باشد. شاخص‌های ارزیابی بسیار متنوعی پیشنهاد شده‌اند که در این پژوهش از شاخص دیویس بولدین جهت ارزیابی خوشه‌ها استفاده شده است [۱۲].

این شاخص از معیار شباهت بین دو خوشه (R_{ij}) استفاده می‌کند که براساس پراکندگی یک خوشه (S_i) و عدم شباهت بین دو خوشه (d_{ij}) تعریف می‌شود. شباهت بین دو خوشه را می‌توان به صورت‌های مختلفی تعریف کرد ولی باید شرایط زیر را دارا باشد:

- $R_{ij} \geq 0$
- Rij=Rji
- اگر si و sj هر دو برابر صفر باشند، آن گاه Rij نیز برابر صفر باشد.
- اگر sj<Sk و dij = dik، آن گاه Rij>Rik
- اگر Sj=Sk، آن گاه dij<dik، آن گاه Rij >Rik

معمولاً شباهت بین دو خوشه به صورت زیر تعریف می‌شود:

فرمول ۲ $$Rij = \frac{sj+si}{dij}$$

که در آن dij و si با روابط زیر محاسبه می‌شوند:

$$Si=\frac{1}{|ci|}\sum_{x\in ci}^{\infty} d(x, vi) \quad = d(vi,vj) \quad Dij$$

با توجه به مطالب بیان شده و تعریف شباهت بین دوخوش، دیویس بولدین به صورت زیر تعریف می‌شود:

$$DB=\frac{1}{nc}\sum_{i=1}^{nc} Ri$$

که Ri در آن به صورت زیر محاسبه می‌شود:

$$Ri=max(Rij) \ , \ i=1\ldots nc$$

این شاخص در واقع میانگین شباهت بین هر خوشه با شبیه‌ترین خوشه به آن را محاسبه می‌کند. هرچه مقدار این شاخص کمتر باشد،خوشه‌های بهتری تولید شده است [۱۲].

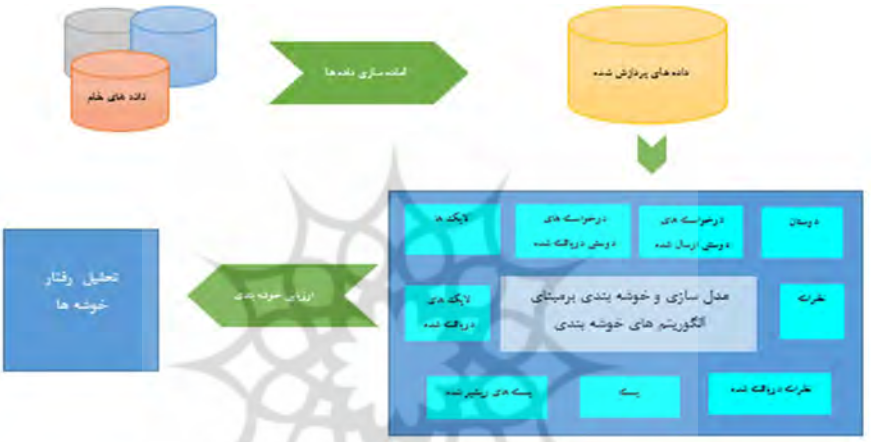
۲-۲- پیشینهٔ تجربی

در سال‌های گذشته شبکه‌های اجتماعی و نقش آنها در زندگی افراد افزایش یافته است. یافته‌های پژوهش‌های انجام شده در این زمینه در جدول زیر آمده است.

عنوان جدول و شماره؟؟؟

محقق (سال)	موضوع	جامعه آماری و نمونه	روش گردآوری و تحلیل داده	یافته‌های تحقیق
حمید خبزی(۲۰۱۵) [۱۵]	ارزیابی سرگرمی افراد در شبکه‌های اجتماعی	۱۰۰ کاربر یک شبکه اجتماعی	K میانگین	خوشه‌بندی کاربران به ۴ خوشه: طرفداران بی‌تفاوت،ضعیف، معمولی و متعصب
آلوارو التیگوزا(۲۰۱۴) [۱۶]	پیش‌بینی شخصیت افراد با کاوش در تعالات اجتماعی در فیسبوک	۲۰۰۰۰ کاربر یک شبکه اجتماعی	نایو بیز نزدیک‌ترین همسایگی درخت تصمیم قوانین انجمنی	الگوریتم‌ها با مدل‌های ۳-کلاس و ۵-کلاس به کار گرفته شده‌اند.
ویلن ون و میشل ولدن [۱۷] (۲۰۱۴)	خوشه‌بندی کاربران	کاربران فیسبوک	کاهش بعد K میانگین MCA	تشخیص چهار خوشه کاربر
احمد حوالا (۲۰۱۵) [۱۸]	پروفایل پویا برای شخصی‌سازی وب‌سایت	–	الگوریتم‌های GEW 3C	ارائه سیستم جستجو شخصی‌سازی شده
رودولا تیسیتسو [۱۹] (۲۰۱۵)	وفاداری به شبکه اجتماعی	۳۲۰ عضو شبکه‌های اجتماعی مختلف	پرسشنامه روش‌های آماری	ارائه مفاهیم نظری و عملی
جیانگ [۱۳] (۲۰۱۴)	خوشه‌بندی کاربران شبکه اجتماعی بر مبنای رفتار احساسی	49 556 کاربر یک شبکه اجتماعی چینی	PCA- شباهت -شباهت فاصله‌ای	این پژوهش تحلیل رفتار احساسی چند متغیره از کاربران شبکه اجتماعی ارائه داده است که امکان خوشه‌بندی کاربران را فراهم می‌کند.
ژائو [۲۰] (۲۰۱۱)	روش جدید خوشه‌بندی در شبکه‌های اجتماعی	–	درخت تصمیم یا گراف	ارائه مدلی برای شناسایی گروه‌ها در شبکه اجتماعی

۳- مدل مفهومی: در این پژوهش هدف این است که پس از پیش‌پردازش داده‌ها و شناسایی داده‌های پرت با استفاده از تکنیک‌های داده‌کاوی، یک الگوی داده‌ای از رفتارهای کاربران شبکه‌های اجتماعی طراحی شود. از طرفی سعی در بررسی این امر است که چرا داده‌ای به گروه خاصی تعلق گرفته است تا با تحلیل آن بتوان در مورد گرایش‌های رفتاری افراد مختلف تصمیم‌گیری کرد. مدل مفهومی پژوهش در شکل ۱ آمده است.



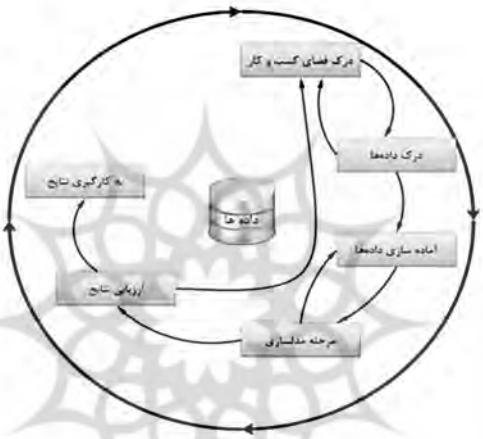
شکل ۱ مدل مفهومی پژوهش

۴- روش‌شناسی پژوهش

این پژوهش، از نظر هدف کاربردی است و از نظر نحوه گردآوری داده‌ها در گروه مطالعات موردی قرار می‌گیرد. در این پژوهش کاربرانی که از شبکه اجتماعی مورد نظر استفاده می‌کنند که شامل تعداد ۳۱۰۳۳ کاربر می‌باشد، جامعه آماری در نظر گرفته شده‌اند. با توجه به عدم توانایی مدل‌های آماری درطراحی مدل‌های پیچیده تحلیل روند و پیش‌بینی و عدم توانایی انسان در شناسایی روابط غیرخطی میان داده‌ها و وجود روابط پنهان میان آنها، پژوهش با به‌کارگیری تکنیک‌های داده‌کاوی و نرم‌افزارهای مربوطه صورت می‌گیرد.

در این پژوهش از مدل فرایندی اجرای پروژه‌های داده‌کاوی برای اجرای آن استفاده شده است که از شش مرحله تشکیل شده است : مرحله درک فضای کسب‌و کار، مرحله درک داده‌ها، مرحله آماده‌سازی داده‌ها، مرحله مدلسازی، مرحله ارزیابی، به‌کارگیری داده‌کاوی.

مدل اجرایی کریسپ یک مدل حلقه‌ای و تکرارشونده است، به این معنا که برخی از مراحل، پس از اجرا ممکن است چندین بار اجرا شوند تا نتیجه موردنظر در مدلسازی حاصل شود. در شکل ۲ مدل فرایندی کریسپ نشان داده شده است [۱۴].



شکل ۲ مدل فرایندی داده‌کاوی

۴-۱- مرحله اول: درک فضای کسب‌وکار

این مرحله شامل شناخت اهداف کسب و کار و تعیین اهداف مورد انتظار از پروژه، **ارزیابی شرایط** و بررسی تمامی فرضیه‌ها، محدودیت‌ها و منابع و**درک اهداف داده‌کاوی** می‌باشد.

۴-۲- مرحله دوم: درک داده‌ها

این مرحله شامل جمع‌آوری داده‌های اولیه، تشریح داده‌ها، اکتشاف در داده‌ها و انجام برخی تجزیه و تحلیل‌های آماری بر داده‌ها، **بررسی کیفیت داده‌ها** و بررسی وجود داده‌های دورافتاده و یا داده‌های از دست رفته را در مجموعه داده می‌باشد.

۴-۳- مرحله سوم: آماده‌سازی داده‌ها

این مرحله شامل انتخاب داده‌ها ،پاک‌سازی داده‌ها، تولید داده و ایجاد ویژگی می‌باشد.

۴-۴- مرحله چهارم: مدلسازی

در این مرحله، الگوریتم‌های مناسب داده‌کاوی انتخاب شده و با توجه به هدف تحقیق، بر داده‌ها اعمال می‌شوند.

۴-۵- مرحله پنجم

ارزیابی در این گام براساس الگوریتم‌های پیاده‌سازی شده روی داده‌ها، میزان اعتبار مدل خروجی با استفاده از داده‌های جدید و ارزیابی مدل بر داده‌های فعلی صورت می‌پذیرد. در صورت تأیید مدل و عدم نیاز به پیاده‌سازی مجدد مدل، گام ششم اجرا می‌شود.

۴-۶- مرحله ششم: به‌کارگیری

پس از ارزیابی مدل، باید برنامه‌ای برای به‌کارگیری مدل در دنیای واقعی تنظیم شود. در این برنامه باید به منظور نگهداری و نظارت از مدل داده‌کاوی پیش بینی‌های لازم انجام شود. داده‌ها در هر سازمانی پیوسته در حال تغییر هستند و مدل ساخته شده هر چند وقت یک‌بار باید بر این داده‌ها اعمال شود تا سازمان بتواند با به‌کارگیری مدل‌های ساخته شده با دقت بیشتری تصمیم‌های مقتضی را اعمال کند.

هدف از این پژوهش، خوشه‌بندی کاربران شبکه اجتماعی و توصیف هریک از خوشه‌ها براساس رفتار کاربر در شبکه می‌باشد، به عبارت دیگر در این مقاله کاربران شبکه اجتماعی براساس اینکه در طول مدت عضویت خود در شبکه چه فعالیت‌هایی انجام داده‌اند، خوشه‌بندی شده و سپس رفتار آنها بررسی شده است. در این راستا از الگوریتم خوشه‌بندی K-Means استفاده شده است که نسبت به سایر روش‌ها، از دقت بالاتری در خوشه‌بندی برخوردار است. مجموعه داده‌های مورد استفاده برای این مقاله کاربران یکی از شبکه‌های اجتماعی- ایرانی پرکاربرد است که شامل تعداد ۳۱۰۳۳ کاربر فعال بوده و اقلام اطلاعاتی بر مبنای ابعاد مختلف فعالیت آنها گردآوری شده است.

در مرحله پیش‌پردازش داده‌ها، هدف شناسایی و جدا کردن کاربرانی بود که از بین کاربران، مناسب با هدف و مسئله این تحقیق بودند. منظور از "مناسب"، این بوده است که کاربر به فعالیت در شبکه اجتماعی علاقه کافی داشته باشد. پیش‌پردازش داده‌ها در این پژوهش در چند مرحله انجام شده است:

نخست تعدادی از کاربرانی که مناسب هدف پژوهش نبودند، حذف گردیدند. این کاربران به شرح زیر پالایش شدند:

• کاربرانی که تنها در شبکه مورد نظر عضو بودند و هیچ‌گونه فعالیتی نداشتند، از بین کاربران مجموعه داده حذف شدند. این کاربران تعداد ۱۰۴ نفر از کل کاربران را تشکیل می‌دهند؛

• کاربرانی که تنها به فعالیت دوست‌یابی در شبکه پرداختند و هیچ‌گونه فعالیت دیگری نیز در شبکه نداشتند، از بین مجموعه داده‌ها حذف شدند. تعداد کاربرانی که در این مرحله حذف شدند، ۱۴۸۶۱نفر می‌باشند؛

• کاربرانی که بجز فعالیت دوست‌یابی و پست هیچ‌گونه فعالیت دیگری نداشتند و یا بازخوردی از سایر اعضا نسبت به پست‌های خود دریافت نکرده‌اند، نیز از بین مجموعه داده اصلی حذف شد. تعداد کاربران حذف شده ۵۰۷۱ نفر از بین کل کاربران می‌باشد.

پس از بررسی داده‌های کاربرانی که جهت تحلیل انتخاب شده‌اند، نیاز به متغیری بود که قبل از خوشه‌بندی کاربران را براساس آن تقسیم‌بندی کرده تا بتوان کاربرانی که تا حدی از لحاظ میزان فعالیت به هم شباهت دارند، وارد الگوریتم خوشه‌بندی شوند. با کمک متغیر جدید می‌توان امتیازی به هر کاربر اختصاص داد. متغیر جدید برحسب سایر متغیرها و به‌صورت مجموع فعالیت‌های فرد در طول مدت عضویت خود در شبکه ایجاد شده است. کاربران براساس این متغیر جدید به چهار دسته کلی تقسیم می‌شوند.

• کاربران غیرفعال: کاربرانی که امتیاز کسب شده آنها کمتر از ۱۰۰ می‌باشد. تعداد کاربران این دسته ۶۶۰۵ نفر از کل کاربران می‌باشد.

• کاربران منفعل: کاربرانی که امتیاز کسب شده آنها بین ۱۰۰ تا ۱۰۰۰ می‌باشد. تعداد کاربران این دسته ۲۵۷۹ نفر از کل کاربران می‌باشد.

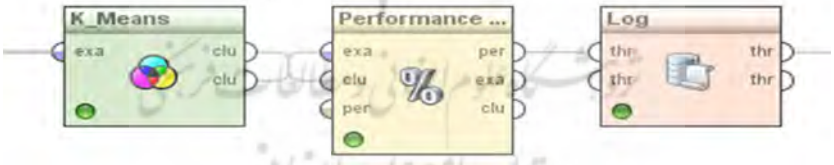
• کاربران فعال: کاربرانی که امتیاز کسب شده آنها بین ۱۰۰۰ تا ۳۰۰۰ می‌باشد. تعداد کاربران این دسته ۹۰۶ نفر از کل کاربران می‌باشد.

• کاربران بسیار فعال: کاربرانی که امتیاز کسب شده آنها بیش از ۳۲۰۰ می‌باشد. تعداد کاربران این دسته نفر ۹۰۷ از کل کاربران می‌باشد.

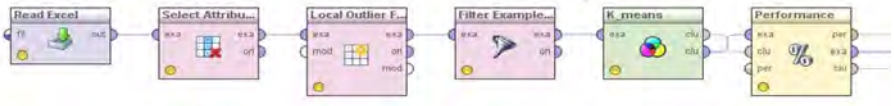
درنهایت برای هر دسته از کاربران جهت تشخیص داده‌های پرت از شاخص عامل پرت محلی استفاده شده است تا بتوان با انجام پیش‌پردازش روی داده‌ها کیفیت نتیجه داده‌کاوی را بهبود بخشید (شکل ۳). عملیات خوشه‌بندی کاربران برای هرکدام از این دسته از کاربران به صورت جداگانه انجام شده است.

شکل ۳ مراحل انجام پیش‌پردازش روی داده‌ها

قبل از اجرای الگوریتم خوشه‌بندی K میانگین، لازم است که تعداد خوشه‌های مناسب برای هر کدام از دسته‌های کابران شناسایی شود. در این مقاله برای دست‌یابی به تعداد مناسبی از خوشه در هر دسته الگوریتم مورد نظر برای هشت بار با تعداد مختلف خوشه (K=3,4,…,10) با اعمال سایر پارامترها از جمله نوع مقیاس «واگرایی برگمن» و همچنین نوع واگرایی «فاصله اقلیدسی»اجرا شد. مراحل اجرای پیش‌پردازش و تعیین تعداد خوشه مناسب و درنهایت خوشه‌بندی کاربران در نرم‌افزار رپیدماینر در شکل‌های ۴ و ۵ نشان داده شده است.

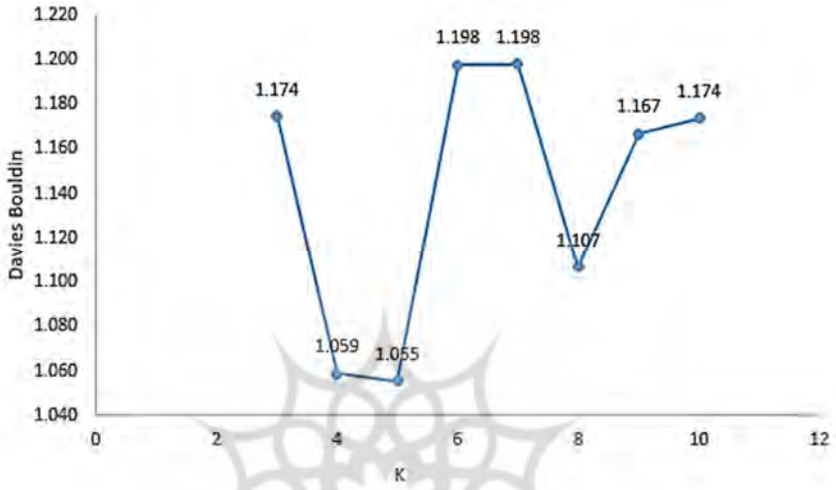


شکل ۴ مراحل تعیین تعداد خوشه بهینه

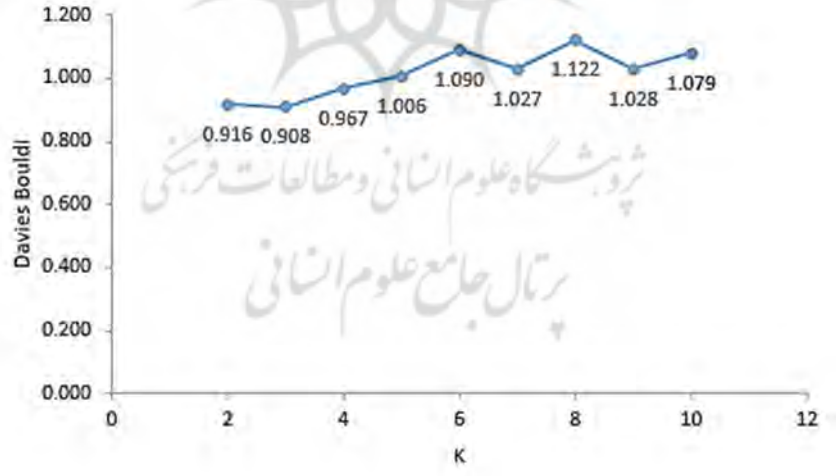


شکل ۵ مراحل انجام مدلسازی روی داده‌ها با استفاده از تعداد خوشه بهینه به دست آمده برای هر دسته از کاربران

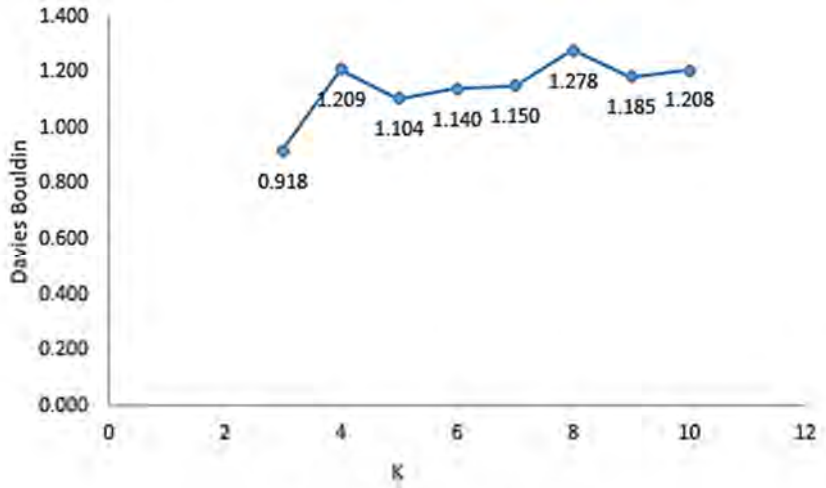
شکل های ۶ الی ۹، نتیجه ارزیابی مدلسازی‌های انجام شده با پارامترهای تعیین شـده را بـا استفاده از محاسبه شاخص دیویس بولدین نشان می‌دهد.



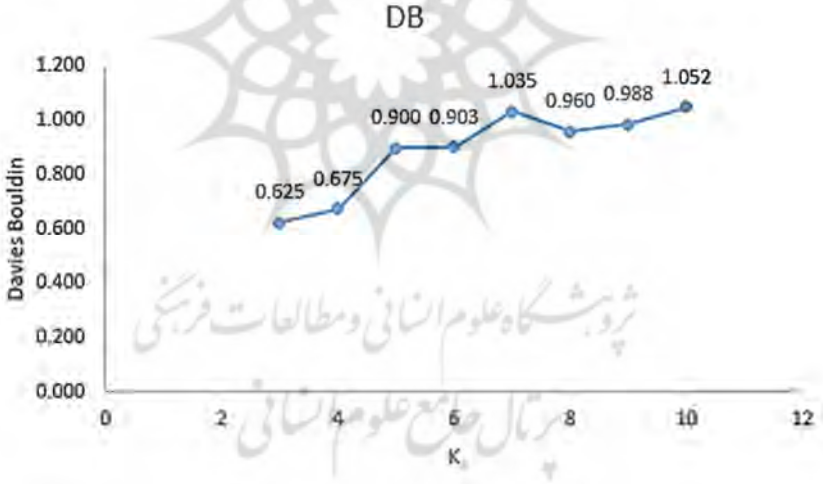
شکل ۶ کاربران غیرفعال



شکل ۷ کاربران منفعل



شکل ۸ کاربران فعال



شکل ۹ کاربران بسیار فعال

همان طور که در بخش گذشته گفته شد، هرچه شاخص دیویس بولدین کمتر باشد، مدلسازی انجام شده بهینه‌تر بوده است .از این رو تعداد خوشه ۵ برای کاربران غیرفعال و تعداد خوشه ۳ برای سایر کاربران که دیویس بولدین آنها از همه کمتر شده است، از سایر تعداد خوشه‌ها نتیجه بهتری داشتند. درنهایت، کاربران هر دسته به تعداد خوشه بهینه با ویژگی‌های متفاوت تقسیم شدند.

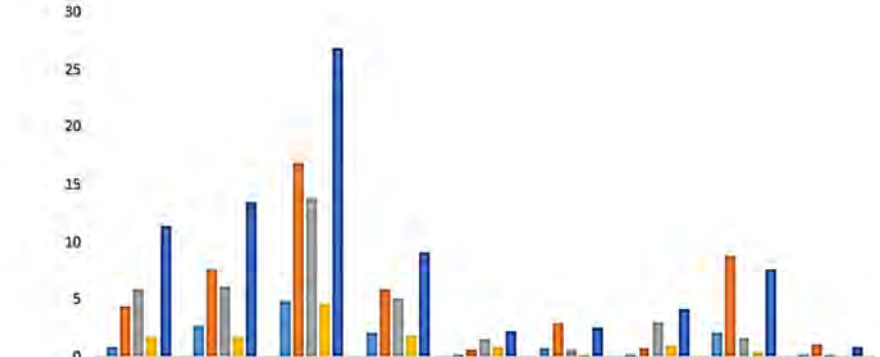
۵- یافته‌های پژوهش

۵-۱- تحلیل خوشه‌ها

در این قسمت به توصیف کاربران هر دسته و خوشه‌های آنها براساس ویژگی‌های آنها می‌پردازیم.

۵-۱-۱- کاربران غیرفعال

این دسته از کاربران به پنج خوشه تقسیم شده‌اند و حجم زیادی از کاربران در دو خوشه یک و چهار قرار گرفته‌اند بجز خوشه دوم که بیشتر از زنان مجرد تشکیل شده است، اکثریت کاربران این دسته مردان مجرد می‌باشند که دارای سن حدود ۲۹ سال و تحصیلات دیپلم هستند. بررسی فعالیت‌های کاربران این دسته نشان از حجم بالای درخواست‌های دوستی دریافت شده به‌وسیله این کاربران نسبت به تعداد درخواست‌های دوستی ارسال شده آنها بیشتر می‌باشد. درواقع این دسته از کاربران مورد توجه سایر اعضای شبکه جهت ارتقای ارتباطات در شبکه می‌باشد. با این حال تعداد دوستان این دسته کاربران نشان‌دهنده این است که آنها توجه چندانی به درخواست‌های دوستی دریافت شده‌شان ندارند و علاقه دارند تا با کسانی که می‌شناسند در شبکه ارتباط داشته باشند. خوشه پنجم فعال‌ترین کاربران این دسته می‌باشند، کاربران خوشه دوم مدت زمان کمی است که در شبکه عضو بوده‌اند و کاربران خوشه اول نسبت به پست‌ها و نظرات سایر اعضای شبکه هیچ واکنشی نشان نداده‌اند و بی‌تفاوت هستند. جزئیات مربوط به میانگین فعالیت‌های کاربران در این دسته در نمودار ۱ نشان داده شده است.



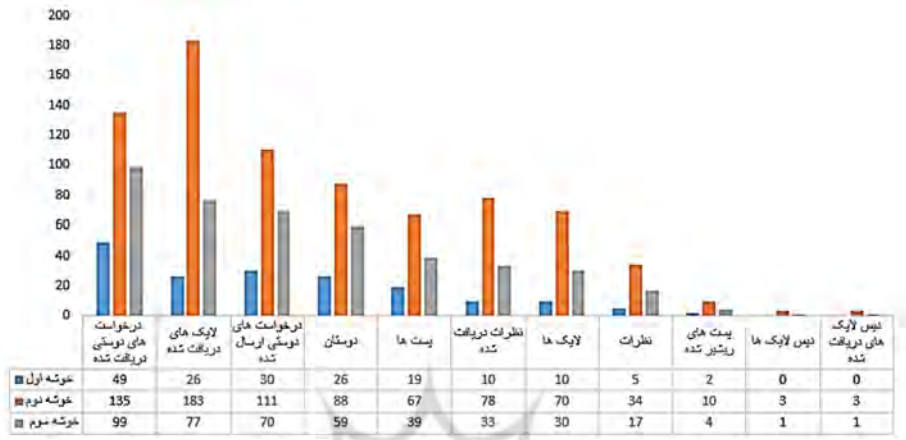
	دوستان	درخواست های دوستی ارسال شده	درخواست های دوستی دریافت شده	پست ها	نظرات	نظرات دریافت شده	لایک ها	لایک های دریافت شده	پست های ریتیوت شده
خوشه اول	1	3	5	2	0	1	0	2	0
خوشه دوم	4	8	17	6	1	3	1	9	1
خوشه سوم	6	6	14	5	2	1	3	2	0
خوشه چهارم	2	2	5	2	1	0	1	0	0
خوشه پنجم	11	14	27	9	2	3	4	8	1

نمودار ۱ جزئیات مربوط به میانگین فعالیت دسته کاربران غیرفعال

۵-۱-۲- کاربران منفعل

کاربران این دسته به سه خوشه تقسیم شده‌اند، خوشه اول بیشترین تعداد اعضا و خوشه سوم کمترین تعداد اعضا را دارا می‌باشند. کاربران این دسته بیشتر از مردان مجرد تشکیل شده است اما خوشه دوم بیشتر از زنان مجرد تشکیل شده است، مدرک تحصیلی لیسانس و شغل کارمند بیشترین تعداد را در این دسته داشته است. خوشه دوم دارای جوان‌ترین کاربران و جدیدترین کاربران این دسته می‌باشد. بررسی رفتارهای کاربران این دسته در شبکه مذکور نشان‌دهنده بالابودن تعداد لایک‌های دریافتی کابران این دسته و همچنین بالابودن مقدار متغیرهای مربوط به فعالیت‌های دوست‌یابی در این دسته می‌باشد. تعداد درخواست‌های دوستی دریافت شده به‌وسیله این کاربران نسبت به تعداد درخواست‌های دوستی ارسال شده آنها بیشتر می‌باشد که نشان‌دهنده محبوبیت این کاربران در بین کاربران شبکه اجتماعی می‌باشد. تمرکز کاربران خوشه اول بر فعالیت دوست‌یابی در شبکه می‌باشد. بیشترین تعداد پست در بین کاربران این دسته متعلق به کاربران خوشه دوم می‌باشد در عین حال تعداد لایک‌های دریافتی کاربران این

خوشه نیز بالا بوده است. جزئیات مربوط به میانگین فعالیت‌های کاربران در این دسته در نمودار ۲ نشان داده شده است.



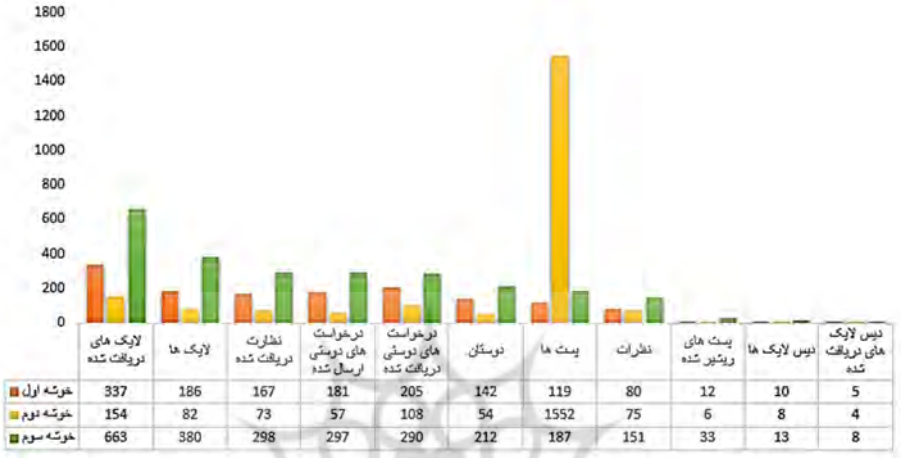
	درخواست های دوستی دریافت شده	لایک های دریافت شده	درخواست های دوستی ارسال شده	دوستان	پست ها	نظرات دریافت شده	لایک ها	نظرات	پست های ریشور شده	دیس لایک ها	دیس لایک های دریافت شده
خوشه اول	49	26	30	26	19	10	10	5	2	0	0
خوشه دوم	135	183	111	88	67	78	70	34	10	3	3
خوشه سوم	99	77	70	59	39	33	30	17	4	1	1

نمودار ۲ جزئیات مربوط به میانگین فعالیت دسته کاربران منفعل

۵-۱-۳- کاربران فعال

کاربران این دسته به سه خوشه تقسیم شده‌اند که خوشه اول بیشترین تعداد اعضا و خوشه دوم کمترین تعداد اعضا را دارا می‌باشند. ترکیب کاربران خوشه دوم و سوم بیشتر از مردان مجرد تشکیل شده است در حالی که خوشه اول بیشتر از زنان مجرد تشکیل شده است. خوشه اول دارای بیشترین تعداد کاربر و مسن‌ترین کاربران می‌باشد. مدرک تحصیلی کاربران خوشه اول و سوم لیسانس بوده در حالی که مدرک تحصیلی کاربران خوشه دوم دیپلم بوده است. کاربران خوشه اول طلبه هستند. کاربران خوشه دوم دارای شغل آزاد می‌باشند و کاربران خوشه سوم دانشجو هستند. کاربران خوشه دوم از اعضای قدیمی شبکه هستند. مشاهده رفتار کاربران این دسته و بازخوردهایی که از سایر کاربران شبکه دریافت کرده‌اند، نشان‌دهنده بالا بودن تعداد لایک‌های دریافتی، تعداد لایک‌ها و تعداد کامنت‌های دریافتی این دسته می‌باشد. تعداد درخواست‌های ارسالی این دسته از کاربران بیش از تعداد درخواست‌های دریافتی آنان می‌باشد. تعداد پست‌های کاربران خوشه دوم بسیار بالا بوده است در عین حال بازخورد دریافتی از سایر کاربران نسبت به پست‌های این دسته بسیار پایین است. تعداد لایک‌ها و

نظرات دریافت شده کاربران خوشه بالا بوده و نشان‌دهنده محبوبیت پست‌های این کاربران در نظر سایر کاربران می‌باشد.جزئیات مربوط به میانگین فعالیت‌های کاربران در این دسته در نمودار ۳ نشان داده شده است.



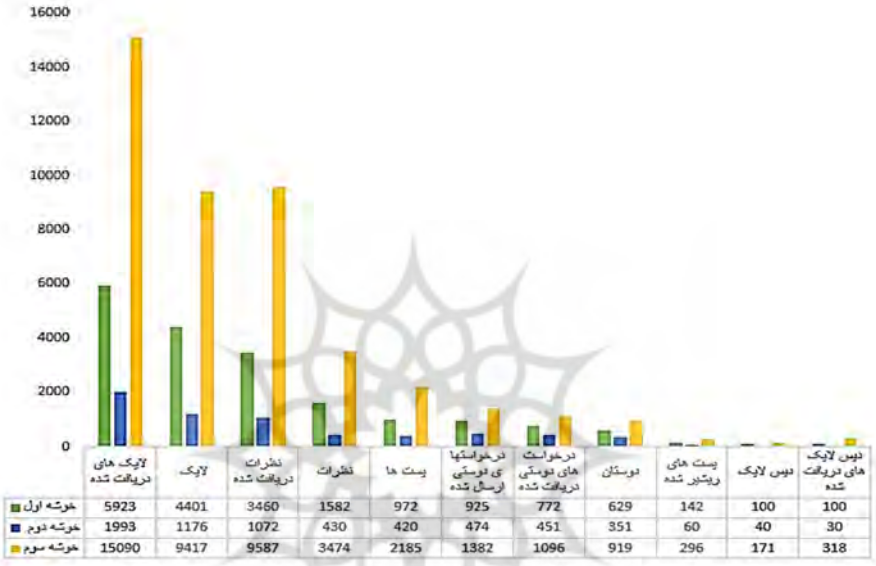
	لایک های دریافت شده	لایک ها	نظارت دریافت شده	درخواست های دوستی ارسال شده	درخواست های دوستی دریافت شده	دوستان	پست ها	نظارات	پست های ویرایش شده	دیس لایک ها	دیس لایک های دریافت شده
خوشه اول	337	186	167	181	205	142	119	80	12	10	5
خوشه دوم	154	82	73	57	108	54	1552	75	6	8	4
خوشه سوم	663	380	298	297	290	212	187	151	33	13	8

نمودار ۳ جزئیات مربوط به میانگین فعالیت دسته کاربران فعال

۵-۱-۴- کاربران بسیار فعال

کاربراین این دسته نیز به سه خوشه تقسیم شده‌اند که خوشه دوم بیشترین تعداد اعضا و خوشه سوم کمترین تعداد اعضا را دارا می‌باشند. جمعیت زنان و مردان در این دسته از کاربران تقریباً برابر می‌باشد. همچنین تعداد قابل توجهی از کاربران این دسته مجرد هستند. ترکیب کاربران خوشه اول و سوم بیشتر از مردان مجرد تشکیل شده است در حالی که خوشه اول بیشتر از کاربران هر دو جنسیت مجرد تشکیل شده است. خوشه سوم دارای کمترین تعداد کاربر و مسن‌ترین کاربران نسبت به سایر خوشه‌ها می‌باشد. حجم زیادی از کاربران این دسته دارای مدرک تحصیلی لیسانس می‌باشند. دانشجویان در تمام خوشه‌ها بیشترین تعداد را دارا می‌باشند. با مشاهده رفتار این دسته از کاربران در شبکه اجتماعی آشکار می‌شود که تعداد لایک‌های دریافتی این دسته از کاربران نیز بسیار بالا بوده و نشان‌دهنده این می‌باشد که پست‌های این دسته از کاربران به‌شدت مورد توجه سایر کاربران شبکه قرار می‌گیرد. تعداد لایک‌ها و تعداد کامنت‌های دریافت‌شده این دسته از کاربران نیز بالا می‌باشد. به‌طور میانگین

تعداد درخواست های دوستی ارسال شده این افراد نسبت به درخواست های دوستی دریافت شده آنها بیشتر می‌باشد. پست‌های کاربران این دسته جزء پست‌های داغ شبکه قرار می‌گیرد خوشه سوم دارای بیشترین تعداد پست داغ و خوشه دوم کمترین تعداد پست داغ می‌باشند. جزئیات مربوط به میانگین فعالیت‌های کاربران در این دسته در نمودار ۴ نشان داده شده است.



نمودار ۴ جزئیات مربوط به میانگین فعالیت دست کاربران بسیار فعال

۶- نتیجه‌گیری

در این پژوهش به خوشه‌بندی کاربران شبکه اجتماعی و توصیف هریک از خوشه‌ها براساس رفتار کاربر در شبکه پرداخته شده است. به عبارت دیگر، کاربران شبکه اجتماعی را براساس اینکه در طول مدت عضویت خود در شبکه چه فعالیت‌هایی انجام داده اند، خوشه‌بندی کرده و سپس رفتار آنها بررسی و تحلیل شده است. با بررسی سایر پژوهش‌های انجام شده، مشخص گردید که محققان تنها به شناسایی و تحلیل برخی از متغیرهای مربوط به فعالیت افراد در شبکه اجتماعی پرداخته‌اند و پژوهش مجزایی که به بررسی و تحلیل رفتار خوشه‌های کاربران در یک شبکه اجتماعی کاملاً ایرانی پرداخته باشد، یافت نگردید. وجه تمایز پژوهش

حاضر از سایر پژوهش‌های مرتبطی که تاکنون صورت گرفته است، در استفاده مجموعه کاملی از متغیرهای مربوط به فعالیت افراد در شبکه اجتماعی و خوشه‌بندی حالت‌های مختلف رفتاری کاربران می‌باشد. همچنین تفاوت ویژه این پژوهش نسبت به سایر مطالعات پیشین، بررسی رفتار کاربران ایرانی می‌باشد. جهت تشخیص داده‌های پرت از شاخص عامل پرت محلی استفاده شده است تا بتوان با انجام پیش‌پردازش بر روی داده‌ها، کیفیت نتیجه داده‌کاوی را بهبود بخشید.

با توجه به عدم توانایی مدل‌های آماری در طراحی مدل‌های پیچیده تحلیل روند و پیش‌بینی و عدم توانایی انسان در شناسایی روابط غیرخطی میان داده‌ها و وجود روابط پنهان و منظم میان آنها، پژوهش با به کارگیری تکنیک‌های داده‌کاوی و نرم‌افزارهای مربوطه، به یافتن این الگوها پرداخته است. در این پژوهش از مدل فرایندی اجرای پروژه‌های داده‌کاوی برای اجرای تحلیل استفاده شد و شش مرحله آن شامل درک فضای کسب‌وکار، درک داده‌ها، آماده سازی داده‌ها، مدلسازی، ارزیابی و تحلیل نتایج به طور کامل پوشش داده شدند.

در مرحله پیش‌پردازش داده‌ها، هدف شناسایی و جداکردن کاربرانی بود که از بین کاربران ، مناسب با هدف و مسئله این تحقیق بودند. پس از بررسی داده‌های کاربرانی که جهت تحلیل انتخاب شده‌اند، نیاز به متغیری بود که قبل از خوشه‌بندی، کاربران را براساس آن تقسیم‌بندی کرده تا بتوان کاربرانی که تا حدی از لحاظ میزان فعالیت به هم شباهت دارند، خوشه‌بندی گردند. با کمک متغیر جدید می‌توان امتیازی را نیز به هر کاربر اختصاص داد. متغیر جدید بر حسب سایر متغیرها و به صورت مجموع فعالیت‌های فرد در طول مدت عضویت خود در شبکه ایجاد شده است. در نهایت، کاربران هر دسته به تعداد خوشه بهینه با ویژگی‌های متفاوت تقسیم شدند. در این راستا از الگوریتم خوشه‌بندی K میانگین استفاده شده است که نسبت به سایر روش‌ها، دقت بالاتر و خوشه‌بندی بهتری ارائه کرده است. برای تعیین تعداد خوشه بهینه برای هر دسته از کاربران از شاخص دیویس بولدین با اعمال سایر پارامترها از جمله نوع مقیاس «واگرایی برگمن» و همچنین نوع واگرایی«فاصله اقلیدسی» استفاده شده است. درنهایت مشخص شد که کاربران براساس نیازها و علایق خود، میزان فعالیت در شبکه‌های اجتماعی را تنظیم می‌کنند و در صورتی که کارکردها و امکانات شبکه اجتماعی، به نیازها و سلایق عمومی

و مشترک کاربران نزدیک‌تر باشد، امکان افزایش تعداد کاربران و همچنین بهبود سطح فعالیت آنها به میزان قابل توجهی وجود دارد.

۷- منابع

[1] R.P. Bagozzi, U.M. Dholakia (2006) "Open source software user communities: A study of participation in Linux user groups", *Management Science*, 52 (7):1099–1115.

[2] H.U. Buhl (2008) "Online communities", Wirtschaftsinformatik 50 (2):81–84.

[3] M. Gneiser J., Heidemann M., Klier A., Landherr F., Probst (2012) Valuation of online social networks taking into account users' interconnectedness, Information Systems and e-Business Management, 10 (1):61–84.

[4] C. Dwyer S., Hiltz K. Passerini (2007) Trust and privacy concern within social networking sites: A comparison of facebook and MySpace, in Proceedings of the Americas Conference on Information Systems –AMCIS, (paper 339).

[5] D.M. Boyd, N.B. Ellison (2007) "Social network sites: definition, history, and scholarship", *Journal of Computer-Mediated Communication*, 13 (1):210–230.

[6] D. Beer (2008) Social network(ing) sites. . . revisiting the story so far: A response to Danah Boyd and Nicole Ellison, Journal of Computer- Mediated Communication, 13 (2): 516–529.

[7] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Introduction to Data Mining, 0nd ed., Wiley, 0274.

[8] Chen, M. S., Han, J., & Yu, P. S. (1996) "Data mining: An overview from a database perspective, Knowledge and data Engineering", *IEEE Transactions on*, 8, 6, 866-883.

[9] Breunig MM, Kriegel H-P, Ng RT, Sander J. (2000) LOF: Identifying density-based local outliers. ACM Sigmod Rec; 29(2): 93–104.

[10] Ravneet Kaur, Sarbjeet Singh (2015) "A survey of data mining and social network analysis based anomaly detection techniques", *Egyptian Informatics Journal*.

[11] Hartigan JA. (1975) *Clustering algorithms*, John Wiley & Sons, Inc. New York, NY, USA.

[12] D.L. Davies and D.W. Bouldin (1979) A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2):224{227, 1979}.

[13] ZHU Jiang , WANG Bai, WU Bin, Social network users clustering based on multivariate time series of emotional behavior, April 2014, 21(2): 21–31, www.sciencedirect.com/science/journal/10058885

[14] Chapman P., Clinton J., Kerber, R., Khabaza T. Reinartz, T. Shearer, C.Wirth (2000) 100 step-by-step data mining guide. Technical report, CRISP-DM.

[15] Hamid Khobzi, Babak Teimourpour (2015) "LCP segmentation: A framework for evaluation of user engagement in online social networks", *Computers in Human Behavior*, 50: 101–107.

[16] Alvaro Ortigosa, Rosa M. Carro, José Ignacio Quiroga (2014) "Predicting user personality by mining social interactions in Facebook", *Journal of Computer and System Sciences*, 80:57–71.

[17] Jan-Willem van Dam, Michel van de Velden Online profiling and clustering of Facebook users, Decision Support Systems (2014), doi: 10.1016/j.dss.2014.12.001

[18] Ahmad Hawalah, Maria Fasli, Dynamic user profiles for web personalization, Expert Systems with Applications 42 (2015) 2547–2569.

[19] Rodoula H. Tsiotsou (2015) "The role of social and parasocial relationships on social networking sites loyalty", *Computers in Human Behavior*, 48:401–414.

[20] Peixin Zhao, Cun-Quan Zhang (2011) "A new clustering method and its application in social networks", *Pattern Recognition Letters*, 32: 2109–2118.