4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
چهارمین کنفرانس بین المللی وب پژوهی
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

Web Users Analysis Using Clustering Algorithms

[1] Zohreh Shokrollahi, [2] Asghar Karimi

[1] Lecturer at Institute of Higher Education ACECR Esfahan
z.shokrollahi2012@gmail.com

[2] Instructor at Institute of Higher Education ACECR Esfahan
AS.karimi52@gmail.com

Abstract

With the rapid development of the World Wide Web and increasing the volume of information, Web research has become an important research area. Web mining research is mainly categorized into two types of web content mining and web usage mining. An important topic in web usage mining is the clustering of users in other words, grouping these users into clusters based on their common features. In this paper, using k-means, Kohonen, and TwoStep methods, we clustered the users into groups with similar characteristics and used the principal component analysis method to enhance clustering quality and use the silhouette criterion to assess clustering quality. Among these three methods, k-means with two clusters had the highest quality and the data set was clustered and analyzed using this method. By analyzing the clusters, can get a better understanding of the users and provide custom and more convenient services for them.

Keywords: Web mining, Clustering, K-means algorithm, TwoStep algorithm, Kohonen algorithm.

4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

تحلیل کاربران وب با استفاده از الگوریتم های خوشه‌بندی

زهره شکراللهی[۱]، اصغر کریمی[۲]

[۱] مدرس مؤسسه آموزش عالی جهاددانشگاهی اصفهان
z.shokrollahi2012@gmail.com

[۲] عضو هیأت علمی مؤسسه آموزش عالی جهاددانشگاهی اصفهان
AS.karimi52@gmail.com

چکیده

با توسعه‌ی سریع شبکه جهانی وب و افزایش حجم اطلاعات، وب پژوهی به حوزه تحقیقاتی مهمی تبدیل شد. تحقیقات در وب پژوهی به طور عمده به دو دسته‌ی کاوش محتوای وب و کاوش استفاده از وب دسته‌بندی می‌شود. یک موضوع مهم در کاوش استفاده از وب، خوشه‌بندی کاربران به عبارت دیگر گروه‌بندی این کاربران به خوشه‌هایی بر اساس ویژگی‌های مشترک آن‌ها است. در این مقاله با استفاده از سه روش k-میانگین، کوهونن و خوشه‌بندی دومرحله‌ای کاربران را به گروه‌هایی با ویژگی‌های مشابه خوشه‌بندی کرده و جهت افزایش کیفیت خوشه‌بندی از روش تحلیل مؤلفه اصلی و برای ارزیابی کیفیت خوشه‌بندی از معیار سیلهوت استفاده نمودیم. در بین این سه روش k-میانگین با دو خوشه بالاترین کیفیت را داشته و مجموعه داده با استفاده از این روش خوشه‌بندی و تحلیل شد. نتایج و یافته‌های این تحقیق با بهینه‌سازی خوشه‌بندی می‌تواند برای دانشجویان، طراحان وب و ارائه‌دهندگان خدمات و تبلیغات اینترنتی مفید بوده و به ارائه‌ی خدمات هدفمند و سفارشی به کاربران منجر شود.

کلمات کلیدی

وب پژوهی، خوشه‌بندی، الگوریتم k-میانگین، الگوریتم خوشه‌بندی دومرحله‌ای، الگوریتم کوهونن.

۱- مقدمه

شبکه جهانی وب رسانه‌ای محبوب و تعاملی برای انتشار اطلاعات است. با این شرایط، ما در حال حاضر غرق در اطلاعات بوده و با سربار اطلاعاتی مواجه هستیم [3]. این حجم از اطلاعات کاربران را با مسائل زیر در تعامل با وب مواجه می‌سازد:

۱) پیدا کردن اطلاعات مرتبط: امروزه ابزارهای جستجو دقت پایینی دارند که منجر به نتایج جستجوی بی ارتباط شده و پیدا کردن اطلاعات مرتبط را دشوار می‌سازد.

۲) ایجاد دانش جدید از اطلاعات در دسترس: دشوار بودن ترکیب نتایج یافت شده و استخراج دانش از آن‌ها.

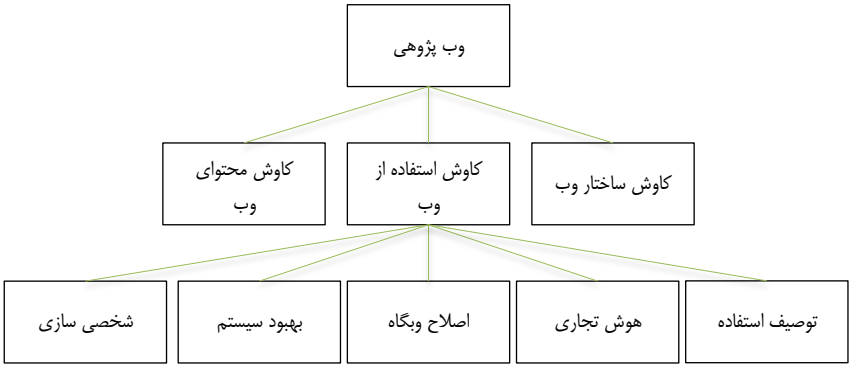
۳) شخصی‌سازی اطلاعات: هنگامی که کاربران با وب تعامل می‌کنند آن‌ها ترجیحات متفاوتی در محتوا و موارد ارائه شده دارند.

۴) یادگیری در مورد کاربران: این مسئله به این موضوع مربوط می‌شود که کاربران چه چیزی می‌خواهند و چه کاری انجام می‌دهند. در کنار این مسئله زیر مسئله‌های دیگری مانند سفارشی‌سازی اطلاعات و طراحی وبگاه و مدیریت تبلیغات برای کاربران وجود دارد.

برای حل مسائل بالا ابزارهای بسیاری مانند پایگاه داده، بازیابی اطلاعات، پردازش زبان طبیعی، داده‌کاوی و غیره وجود دارند [4]. وب پژوهی، استفاده از تکنیک‌های داده‌کاوی برای کشف و استخراج خودکار اطلاعات از اسناد و سرویس‌های وب است. این حوزه‌ی تحقیقاتی به دلیل رشد فزاینده‌ی منابع اطلاعاتی بر روی وب امروزه علاقه‌مندی جوامع تحقیقاتی را به خود جلب کرده است [3]. شکل (۱) دسته‌بندی وب پژوهی به حوزه‌های اصلی را نشان می‌دهد.

از دیدگاه کاربردی و کسب‌وکار، دانش به دست آمده از الگوهای وب می‌تواند به طور مستقیم برای مدیریت مؤثر فعالیت‌های مرتبط با کسب‌وکار الکترونیکی، خدمات الکترونیکی، آموزش الکترونیکی و غیره به کار رود.

اطلاعات دقیق استفاده از وب می‌تواند به جذب مشتریان جدید، حفظ مشتریان فعلی، بهبود بازاریابی و فروش، اثربخشی رشد شرکت‌ها، ردیابی ترک مشتریان و پیدا کردن بهترین ساختار منطقی مؤثر برای فضای وب کمک کند [6]. در این زمینه ارائه اطلاعات و ساختار محیط وب سازگار با سبک شناختی کاربران می‌تواند قابلیت استفاده از سیستم را از لحاظ کارایی و اثربخشی افزایش داده همچنین تجربه‌ی کاربری مثبتی را فراهم نماید



شکل (۱): دسته‌بندی وب پژوهی به حوزه‌های اصلی [5]

[7]. در این بین با استفاده از خوشه‌بندی و گروه‌بندی این کاربران به خوشه‌هایی بر اساس ویژگی‌های مشترک آن‌ها و تجزیه و تحلیل ویژگی‌های گره‌ها، می‌توان به درک بهتری از کاربران دست یافته و خدمات سفارشی و مناسب‌تری فراهم نمود [8]. فرایند خوشه‌بندی گام مهمی در ایجاد پروفایل کاربر بوده و شامل مطالعه‌ی ویژگی‌های مهم بازدیدکنندگان وب است [9]. در ماهیت وب، شخصی‌سازی به انتقال محتوای پویا مانند اجزای متنی، لینک‌ها، تبلیغات، پیشنهاد محصولات و غیره اشاره دارد که درخور نیازها یا علایق کاربر خاص یا بخشی از کاربران است [10].

در ادامه در بخش بعدی ادبیات و کارهای مرتبط با این حوزه بیان شده و در بخش ۳ به تشریح روش مورد استفاده پرداخته و درنهایت در بخش ۴ و ۵ به بیان و تحلیل نتایج می‌پردازیم.

۲- مروری بر ادبیات و کارهای مرتبط

در سال ۱۹۹۶ اتزیونی[1] [11] اولین بار اصطلاح وب پژوهی را ابداع نمود. بنا به گفته‌ی اتزیونی، وب پژوهی استفاده از تکنیک‌های داده‌کاوی برای کشف و استخراج اطلاعات از اسناد و خدمات وب است. وب پژوهی به وظایف کشف منابع، انتخاب و پیش‌پردازش داده‌ها، تعمیم، تحلیل، مصورسازی تقسیم می‌شود. در این بین مدل‌سازی کاربر و شخصی‌سازی زیرمجموعه‌ی تعمیم بوده که این وظیفه به کشف الگوهای کلی در وبگاه‌های شخصی و وبگاه‌های دیگر می‌پردازد [4].

مکانیسم مورد استفاده برای مدل‌سازی کاربر می‌تواند بر اساس روش‌های جمع‌آوری اطلاعات صریح یا ضمنی باشد. اطلاعات صریح به طور مستقیم توسط کاربر معمولاً از طریق فرم‌های ثبت‌نام، پرسشنامه‌ها یا ابزارهای روان‌سنجی مخصوص ارائه می‌شود. از سوی دیگر، اطلاعات ضمنی توسط سیستم به صورت خودکار برای ردیابی رفتار کاربر در تعامل با سیستم به دست می‌آید. تحقیقات متعددی در تلاش برای بررسی مؤثرترین اطلاعات برای مدل‌سازی کاربر انجام شده‌اند. بر اساس [12] هنوز روشن نیست که آیا مدل‌های ضمنی دقیق‌تر هستند یا مدل‌های ایجاد شده توسط اطلاعات صریح. با این وجود اطلاعات ضمنی جمع‌آوری شده بر تعامل انسان کامپیوتر یا بارشناختی کاربر تأثیر نمی‌گذارد؛ از طرفی این روش پیچیده‌تر از روش‌های بازخورد صریح کاربر است زیرا در اغلب موارد داده‌های به دست آمده نامشخص، ناقص یا ناهمگن هستند. روش‌های مدل‌سازی کاربر مبتنی بر اطلاعات شخصی ارائه شده توسط کاربران اغلب از طریق فرم‌های ثبت‌نام بوده و داده‌های جمع‌آوری شده معمولاً شامل اطلاعات جمعیت شناختی، علایق یا ترجیحات کاربران هستند. تکنیک‌های مرسوم برای به دست آوردن اطلاعات صریح شامل جعبه‌ها، فهرست‌های کشویی و یا فیلدهای متنی است که در آن کاربران به طور آزادانه نظر خود را بیان می‌کنند. این تکنیک‌ها این مزیت را دارند که فرمت پاسخ‌ها استاندارد هستند [7].

امروزه فرایند مدل‌سازی کاربران وب به تکنیک‌های داده‌کاوی یا کشف دانش به دلیل بالا بودن حجم داده‌های بر روی وب متصل شده است [7]. در بین روش‌های داده‌کاوی خوشه‌بندی روشی کارآمد برای جستجوی الگوهای پنهان بوده که در ریاضیات، آمار و تحلیل عددی ریشه دارد. از دیدگاه عملی خوشه‌بندی نقش مهمی در کاربردهای داده‌کاوی مانند بازیابی اطلاعات و متن‌کاوی، تجزیه و تحلیل وب و مدیریت ارتباط با مشتریان و بسیاری دیگر دارد [13].

خوشه‌بندی یکی از راه‌های شناسایی الگوهای موجود در بین داده های بسیار است. این روش نمونه‌ها را در خوشه‌های مجزا قرار داده به طوری که داده‌های درون یک خوشه به یکدیگر شبیه و با نمونه‌های سایر خوشه‌ها متفاوت هستند. الگوریتم‌های خوشه‌بندی عموماً همه‌ی متغیرها و مشخصه‌ها را در نظر می‌گیرند در حالی که بسیاری از این مشخصه‌ها غیر مرتبط و اضافه بوده و این مشخصه‌ها نه تنها کمکی به خوشه‌بندی نمی‌نمایند بلکه نتایج را هم تحت تأثیر قرار می‌دهند. همچنین خوشه‌بندی داده‌هایی با ابعاد بالا با مشکلاتی از جمله دشواری درک ابعاد بالا، غیر ممکن بودن نمایش همه ابعاد، رشد نمایی تعداد مقادیر هر بعد و غیر ممکن بودن محاسبه آن‌ها روبرو است. به همین دلیل بهبود کارایی و دقت کاوش اطلاعات از داده‌هایی با ابعاد بالا باید در مرحله پیش‌پردازش داده‌ها با استفاده از روش‌هایی مانند تحلیل مؤلفه اصلی انجام شود [14]. در مرجع [15] چهارده روش کاهش بعد مقایسه شدند که نتایج بیانگر برتری روش تحلیل مؤلفه اصلی به دلیل ماهیت پارامتریک در نگاشت داده‌های با ابعاد بالا به فضایی با ابعاد پایین، نداشتن پارامترهای آزادی که نیاز به بهینه‌سازی داشته باشند، پیچیدگی زمانی و محاسباتی پایین نسبت به سایر روش‌های کاهش بعد است. در ادامه به شرح مختصری از مقالات مربوط به این حوزه می‌پردازیم.

در مقاله [16] از خوشه‌بندی جهت ایجاد یک سیستم پیشنهاددهنده استفاده شده است. در این مقاله یک معیار عدم شباهت بر اساس ارتباط بین دسترسی به صفحات و ساختار نحوی URLهای وبگاه‌ها ارائه شده سپس از الگوریتم K–میانه برای خوشه‌بندی کاربران وب استفاده شده است. در ادامه ارزش خوشه‌های تولید شده توسط دو شاخص اعتبارسنجی

خوشه‌ای ارزیابی شده است. نتایج این مقاله نشان می‌دهد که اندازه‌گیری غیر مستقیم نسبت به سایر روش‌های عدم شباهت مستقل در مورد شاخص‌های اعتبار خوشه‌ای برتر است. در سیستم پیشنهاددهنده‌ی دیگری [17] برای پیشنهاد صفحات وب از خوشه‌بندی دوگانه استفاده شده است. در این روش نقاط قوت خوشه‌بندی مبتنی بر تراکم با خوشه‌بندی k-میانگین ترکیب شده و ایده‌ی اصلی آن استفاده از خوشه‌بندی مبتنی بر تراکم جهت شناسایی تعداد خوشه‌ها و مراکز اولیه‌ی هر خوشه است. نتایج تجربی نشان می‌دهد که این روش تنوع و دقت بیشتری در مقایسه با روش‌های پیشرفته‌تر دارد.

در مقاله [18] روشی به نام CLUE برای خوشه‌بندی URLها ارائه شده است. در این روش مفهوم فاصله URL معرفی و از آن برای ساخت خوشه‌های URL با استفاده از الگوریتم DBSCAN استفاده شده است. به طور خلاصه الگوریتم خوشه‌بندی ارائه شده، ابزاری برای به دست آوردن بینش در مورد داده‌های حمل شده در شبکه، با برنامه‌های کاربردی در زمینه‌های امنیتی یا حفظ حریم خصوصی است.

هدف مقاله [19] معرفی یک معیار شباهت مبتنی بر مربع کای‌دو[۲] برای محاسبه شباهت بین نشست‌ها است. یک روش مبتنی بر مربع کای-دو برای محاسبه ارتباط آماری معنی‌دار بین فرکانس‌های مشاهده شده و مورد انتظار از تعداد صفحات بازدید شده و زمان صرف شده یک کاربر طی یک جلسه ارائه شده است. علاوه بر این یک روش خوشه‌بندی سلسله مراتبی[۳] مبتنی بر مربع کای‌دو برای استخراج اطلاعات مفید از وب لاگ‌ها[۴] معرفی شده است. نتایج تجربی با دو فایل لاگ مختلف نشان می‌دهد که اندازه‌گیری شباهت با الگوریتم Chi-HAC به طور قابل ملاحظه‌ای محاسبات بین اشیای داده‌ای در نشست‌های وب را بهبود داده است.

در مقاله [۱] مدلی برای خوشه‌بندی کاربران وب ارائه شده که در آن تابع علاقه کاربر به هر صفحه با استفاده از زمان سپری شده توسط کاربر در آن صفحه برآورد شده و از این طریق داده‌های مجزای زیادی را از فضای نمونه‌ای حذف نموده‌اند. پس از آن برای پیش‌بینی درخواست بعدی کاربر فعال در وبگاه مفهوم جدید شکاف بین زیر دنباله‌ها ارائه و از یک ماشین پیشنهاددهنده که بر مبنای بلندترین زیر دنباله مشترک دو جلسه کاربری کار می‌کند استفاده شده است.

الگوریتم خوشه‌بندی دیگری جهت شناسایی علایق کاربران با استفاده از پروفایل کاربر در سمت سرور و رفتاری که آن‌ها در حین مرور صفحات دارند در [۲] ارائه شده است. در این روش با استفاده از ترکیب الگوریتم بهینه‌سازی ذرات و الگوریتم خوشه‌بندی فازی، صفحات وب بر اساس علایق کاربر خوشه‌بندی شده و می‌تواند به کاربران برای یافتن صفحاتی مرتبط با علایقشان در کمترین زمان و با دقت بالا کمک نماید.

۳- روش تحقیق

روش تحقیق مورد استفاده در این مقاله شامل مراحل زیر است:

- تعریف مسئله
- انتخاب و تحلیل داده‌ها
- آماده‌سازی داده‌ها
- خوشه‌بندی
- تجزیه و تحلیل خوشه‌ها

۳-۱- تعریف مسئله

مسئله مورد بررسی در این مقاله شناسایی گروه‌هایی از کاربران با علایق و رفتار مشابه و تجزیه و تحلیل ویژگی‌های آن‌ها است. ابزاری که داده‌کاوی جهت شناسایی گروه‌هایی با ویژگی‌های مشابه ارائه می‌نماید خوشه‌بندی می‌باشد.

۳-۲- انتخاب و تحلیل داده‌ها

اطلاعات مورد استفاده در این مقاله شامل اطلاعات ۷۲۱۹ کاربر وب می‌باشد. این مجموعه داده [20] توسط دانشکده فناوری جورجیا با استفاده از پرسشنامه [21] ایجاد شده و شامل ۱۰۲ ویژگی در مورد میزان و موارد استفاده از وب مانند شرکت در گروه‌های چت، کوکی‌ها، استفاده از اخبار الکترونیکی، تعداد دفعات استفاده از وب، میزان ساعاتی که برای تفریح در وب سپری کرده‌اند، میزان ساعات استفاده از وب و غیره می‌باشد.

۳-۳- آماده‌سازی داده‌ها

از آنجایی که تعداد مشخصه‌های مورد استفاده در بخش خوشه‌بندی، تعداد بسیاری مشخصه بوده لازم است به منظور افزایش کیفیت خوشه‌بندی از تکنیک‌های کاهش مشخصه استفاده شود. در این مقاله از روش تحلیل مؤلفه اصلی به عنوان یک روش پیش‌پردازشی استفاده شده است.

۳-۳-۱- تحلیل مؤلفه اصلی

تجزیه‌وتحلیل مؤلفه اصلی یکی از محبوب‌ترین و قدیمی‌ترین روش‌های آماری چند متغیره است. اهداف اصلی این روش شامل موارد زیر است:

- استخراج مهم‌ترین اطلاعات از مجموعه داده
- فشرده‌سازی اندازه مجموعه داده با نگهداری اطلاعات مهم
- ساده‌سازی توصیف مجموعه داده
- تحلیل ساختار مشاهدات و متغیرها

برای رسیدن به این اهداف، روش تحلیل مؤلفه اصلی متغیرهای جدید که مؤلفه‌های اصلی نامیده می‌شوند را با ترکیب خطی متغیرهای اصلی محاسبه می‌نماید [22]. ایده اصلی تحلیل مؤلفه اصلی کاهش ابعاد مجموعه داده‌ای شامل تعداد زیادی از متغیرهای مرتبط است تا جایی که تنوع موجود در داده‌ها حفظ شود. در این روش متغیرهای موجود در یک فضای چند حالته هم‌بسته به یک مجموعه از مؤلفه‌های غیرهم‌بسته خلاصه می‌شوند که هر یک از آن‌ها ترکیب خطی متغیرهای اصلی هستند. مؤلفه‌های غیرهم‌بسته به دست آمده مؤلفه‌های اصلی نامیده می‌شوند که از بردارهای ویژه ماتریس کوواریانس یا ماتریس همبستگی متغیرهای اصلی به دست می‌آیند [23].

۳-۴- خوشه‌بندی

هدف خوشه‌بندی کشف و آشکارسازی الگوهای پنهان در داده‌ها بدون داشتن هدفی از پیش تعیین شده است این روش از روش‌های بدون ناظر می‌باشد.

روش‌های بسیاری برای خوشه‌بندی موجود است و این روش‌ها را می‌توان به سه دسته گسترده روش‌های افرازی[۵]، روش‌های سلسله مراتبی

و روش‌های مبتنی بر چگالی[۶] تقسیم نمود. روش‌های افرازی از معیارهای مبتنی بر فاصله برای خوشه‌بندی نمونه‌های مشابه استفاده می‌نماید. روش‌های سلسله مراتبی، داده‌ها را در سطوح مختلف به روش پایین به بالا یا بالا به پایین افراز می‌کنند؛ این نوع خوشه‌بندی برای خلاصه‌سازی و مصورسازی استفاده می‌شود. روش خوشه‌بندی مبتنی بر چگالی می‌تواند داده‌ها را در اشکال متفاوتی مانند S خوشه‌بندی کند. در این روش نقاط داده‌ای در نواحی متراکم به صورت یک خوشه بوده و به وسیله نواحی با چگالی کمتر از سایر خوشه‌ها جدا می‌شوند [24]. در این مقاله ما به خوشه‌بندی داده‌ها با استفاده از سه روش k–میانگین، روش خوشه‌بندی دومرحله‌ای[۷] و کوهونن[۸] می‌پردازیم.

۳-٤-۱- الگوریتم k–میانگین

الگوریتم k–میانگین یکی از مشهورترین روش‌های خوشه‌بندی است. ورودی این الگوریتم n نمونه داده و مقدار k که تعداد خوشه‌های خروجی را مشخص می‌کند است. در ابتدا تعداد k نمونه به صورت تصادفی از میان کل نمونه‌ها انتخاب می‌شوند. این نمونه‌ها به عنوان مرکز هر خوشه انتخاب شده سپس سایر نمونه‌ها با اندازه‌گیری معیارهایی مانند فاصله اقلیدسی[۹] تا مرکز هر خوشه، درون خوشه‌ای که کمترین فاصله اقلیدسی را تا مرکز آن خوشه دارند قرار می‌گیرند. پس از اضافه کردن اعضای جدید به هر خوشه، مرکز خوشه جدید با محاسبه میانگین بین اعضای هر خوشه تعیین می‌شود. فرایند تکراری تخصیص و به‌روزرسانی مراکز خوشه با هدف حداقل سازی مجموع مجذور خطا برای همه نمونه‌ها است [24]. روابط (۲و۱) نحوه محاسبه مجذور خطا را نشان می‌دهند.

$$SSE(C) = \sum_{K=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2 \quad (۱)$$

$$c_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|} \quad (۲)$$

ck: میانگین خوشه k ام

۳-٤-۲- الگوریتم کوهونن

مدل کوهونن [25] نوعی خاص از مدل شبکه‌ی عصبی است که یادگیری بدون ناظر را انجام می‌دهد. آن بردارهای ورودی را دریافت کرده و نوعی خوشه‌بندی فضایی سازمان یافته یا نگاشت ویژگی را برای گروه‌بندی رکوردهای مشابه انجام داده و فضای ورودی را به یک فضای دو بعدی نگاشت می‌کند که روابط نزدیکی بین خوشه‌ها را تقریب می‌زند.

مدل شبکه کوهونن شامل دو لایه از نورون‌ها یا واحدها است: یک لایه ورودی و یک لایه خروجی. لایه ورودی به طور کامل به لایه خروجی متصل است و هر اتصال یک وزن دارد. در یک مدل کوهونن پارامترها به عنوان وزن بین واحدهای ورودی و واحدهای خروجی نمایش داده می‌شوند یا به طور متناوب به عنوان مرکز خوشه‌ای مرتبط با هر واحد خروجی نمایش داده می‌شوند. رکوردهای ورودی به شبکه ارائه شده و مراکز شبکه در حالتی مشابه به حالتی که در ساخت یک مدل k–میانگین استفاده می‌شود به روز می‌شود با این تفاوت که خوشه‌ها به صورت فضایی در یک شبکه دو بعدی مرتب شده و هر رکورد نه تنها واحدی (خوشه‌ای) که به آن اختصاص داده شده است را تحت تأثیر قرار می‌دهد بلکه واحدهای مجاور واحد برنده را نیز تحت تأثیر قرار می‌دهد.

رابطه (۳) فاصله در شبکه کوهونن که به صورت فاصله اقلیدسی بین بردار ورودی کد شده و مرکز خوشه که برای واحد خروجی محاسبه می‌شود را نشان می‌دهد.

$$d_{ij} = \sqrt{\sum_{k} (x_{ik} - w_{jk})^2} \quad (۳)$$

که x_{ik} مقدار kامین فیلد ورودی برای iامین رکورد است و w_{jk} وزن برای kامین فیلد ورودی روی jامین واحد خروجی است.

۳-٤-۳- الگوریتم خوشه‌بندی دومرحله‌ای

روش خوشه‌بندی دومرحله‌ای یک الگوریتم تجزیه و تحلیل خوشه‌ای مقیاس‌پذیر است که برای رسیدگی به مجموعه داده‌های بزرگ طراحی شده است. این روش دو مرحله دارد: ۱) پیش خوشه‌بندی رکوردها به زیر خوشه‌های کوچک‌تر ۲) خوشه‌بندی زیر خوشه‌های حاصل از مرحله پیش خوشه‌بندی به تعداد مطلوبی خوشه. آن می‌تواند همچنین به طور خودکار تعداد خوشه‌ها را انتخاب کند.

مرحله پیش خوشه‌بندی از روش خوشه‌بندی متوالی استفاده می‌کند. آن رکوردهای داده را یکی‌یکی بررسی کرده و در صورتی که رکورد فعلی باید با خوشه‌های تشکیل شده‌ی قبلی ادغام شود یا باید خوشه‌ای را بر اساس معیار فاصله ایجاد کند تصمیم‌گیری می‌کند.

مرحله خوشه‌بندی، زیر خوشه‌های منتج از مرحله پیش خوشه‌بندی را به عنوان ورودی دریافت کرده سپس آن‌ها را به تعداد مورد نظر گروه‌بندی می‌نماید. در این روش از خوشه‌بندی سلسله مراتبی استفاده می‌شود. در خوشه‌بندی سلسله مراتبی خوشه‌ها به طور بازگشتی ادغام می‌شوند تا زمانی که در انتهای فرایند تنها یک خوشه شامل همه‌ی رکوردها باقی بماند. این فرایند با تعریف یک خوشه‌ی شروع برای هر زیر خوشه تولید شده در مرحله پیش خوشه‌بندی شروع می‌شود. سپس تمام خوشه‌ها مقایسه می‌شوند و جفت خوشه‌ها با کمترین فاصله بین آن‌ها انتخاب شده و به یک خوشه‌ی واحد ادغام می‌شوند. پس از ادغام، مجموعه جدید خوشه‌ها مقایسه شده و نزدیک‌ترین جفت ادغام می‌شوند و این فرایند تکرار می‌شود تا تمام خوشه‌ها ادغام شوند.

روش خوشه‌بندی دومرحله‌ای از معیار فاصله احتمال لگاریتمی[۱۰] برای محاسبه‌ی فیلدهای نمادین و محدوده استفاده می‌کند. روابط (۶-۴) فاصله میان خوشه‌های i و j را نشان می‌دهند [26]:

$$d(i,j) = \xi_i + \xi_j - \xi_{<i,j>} \quad (۴)$$

$$\xi_v = -N_v\left(\sum_{k=1}^{K^A} \frac{1}{2}\log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2\right)\right) + \sum_{k=1}^{K^B} \hat{E}_k \quad (۵)$$

$$\hat{E}_{vk} = -\sum_{l=1}^{L_k} \frac{N_{vkl}}{N_v} log \frac{N_{vkl}}{N_v} \quad (۶)$$

K^A: تعداد نوع محدوده فیلدهای ورودی

K^B: تعداد نوع فیلدهای ورودی از نوع سمبلیک

L_k: تعداد دسته برای kامین فیلد سمبلیک

N_v: تعداد رکوردها در خوشه v

N_{vkl}: تعداد رکوردها در خوشه v متعلق به lامین دسته از kامین فیلد سمبلیک

$\hat{\sigma}_k^2$: واریانس برآورد شده از kامین متغیر پیوسته برای همه‌ی رکوردها

$\hat{\sigma}_{vk}^2$: واریانس برآورد شده از kامین متغیر پیوسته برای همه‌ی رکوردها در خوشه v و

$<i,j>$: یک شاخص نمایش خوشه تشکیل شده با ترکیب خوشه‌های i و j

٤- نتایج خوشه‌بندی

در این بخش به بیان و بررسی نتایج خوشه‌بندی می‌پردازیم. نتایج با استفاده از نرم‌افزار IBM SPSS Modeler 14.2 به دست آمده است. برای ارزیابی کیفیت خوشه‌بندی از معیار سیلهوت[11] استفاده نمودیم. این معیار میزان همبستگی درون خوشه‌ای و میزان جدایی بین خوشه‌ای را نشان می‌دهد و هر چه این میزان بالاتر باشد خوشه‌بندی با کیفیت بالاتری انجام شده است [27]. در نهایت با مقایسه سیلهوت در روش‌های خوشه‌بندی مختلف، بهترین روش انتخاب و خوشه‌های حاصل از آن تحلیل می‌شوند.

٤-١- نتایج خوشه‌بندی دومرحله‌ای

مجموعه داده با و بدون استفاده از روش تحلیل مؤلفه اصلی خوشه‌بندی شد. در این روش نیازی به تعیین تعداد خوشه‌ها نبوده و تعداد خوشه‌های بهینه توسط الگوریتم مشخص می‌شود. نتایج حاصل در جدول (۱) نشان داده شده است.

٤-٢- نتایج خوشه‌بندی کوهونن

روش خوشه‌بندی کوهونن نیز مانند روش خوشه‌بندی دومرحله‌ای نیازی به تعیین تعداد خوشه‌ها ندارد. جدول (۲) نتایج حاصل از این روش را نشان می‌دهد.

٤-٣- نتایج خوشه‌بندی k-میانگین

این الگوریتم با این محدودیت روبرو است که باید کاربر تعداد خوشه‌ها را تعیین نماید. در این پژوهش برای رفع این محدودیت از روش تحلیل مؤلفه اصلی با تعداد مؤلفه‌های ۲، ۳، ۴ و ۵ مؤلفه برای کاهش ابعاد داده‌ها استفاده می‌نماییم. از معیار سیلهوت برای ارزیابی کیفیت خوشه‌بندی با تعداد خوشه‌ها و تعداد مؤلفه‌های متفاوت استفاده نموده و در نهایت تعداد بهینه خوشه را مشخص می‌نماییم. با بررسی میزان سیلهوت به دست آمده در جدول (۳) با تعداد خوشه‌ها و تعداد مؤلفه‌های مختلف مشاهده می‌کنیم که با خوشه‌بندی مجموعه داده به دو خوشه به بالاترین میزان سیلهوت نسبت به تعداد خوشه‌های دیگر دست یافتیم. شکل (۳و۲) نتایج حاصل از خوشه‌بندی را نشان می‌دهند.

جدول (۱): نتایج خوشه‌بندی دومرحله‌ای

	تعداد مؤلفه	تعداد خوشه بهینه	سیلهوت
خوشه‌بندی بدون استفاده از روش تحلیل مؤلفه اصلی	–	۲	۰.۱
خوشه‌بندی با استفاده از روش تحلیل مؤلفه اصلی	۲	۳	۰.۴
	۳	۵	۰.۴
	۴	۱۱	۰.۳
	۵	۱۴	۰.۳

جدول (۲): نتایج خوشه‌بندی کوهونن

	تعداد مؤلفه	تعداد خوشه بهینه	سیلهوت
خوشه‌بندی بدون استفاده از روش تحلیل مؤلفه اصلی	–	۸۸	0> سیلهوت
خوشه‌بندی با استفاده از روش تحلیل مؤلفه اصلی	۲	۱۲	۰.۳
	۳	۱۲	۰.۲
	۴	۱۲	۰.۲
	۵	۱۲	۰.۱

جدول (۳): تحلیل کیفیت خوشه‌بندی و انتخاب تعداد خوشه بهینه

	تعداد مؤلفه	دو خوشه	سه خوشه	چهار خوشه	پنج خوشه
خوشه‌بندی بدون استفاده از روش تحلیل مؤلفه اصلی	–	۰.۱	۰.۱	۰.۱	۰.۱
خوشه‌بندی با استفاده از روش تحلیل مؤلفه اصلی	۲	۰.۵	۰.۴	۰.۴	۰.۳
	۳	۰.۴	۰.۳	۰.۳	۰.۴
	۴	۰.۴	۰.۴	۰.۴	۰.۴
	۵	۰.۴	۰.۴	۰.۳	۰.۳

٥- بحث و بررسی

در این بخش به بررسی و تحلیل نتایج حاصل از خوشه‌بندی k-میانگین با دو خوشه می‌پردازیم. به دلیل بالا بودن تعداد ویژگی‌ها، بعضی از ویژگی‌ها بررسی شده‌اند. نتایج حاصل از این خوشه‌بندی می‌تواند برای طراحان وب و ارائه‌دهندگان خدمات اینترنتی جهت ارائه‌ی خدمات مناسب و سفارشی سودمند باشد.

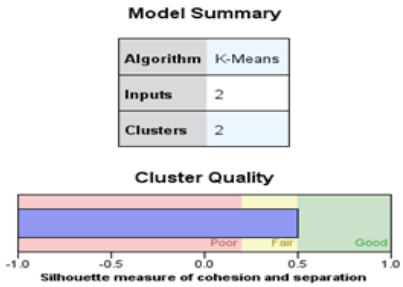
اعضای خوشه دوم بیشترین دفعات استفاده از وب را دارند. شکل (۵و۴) توزیعی از تعداد دفعات استفاده از وب در خوشه‌ها را نشان می‌دهد. ارائه‌دهندگان سرویس‌های اینترنتی می‌توانند از این ویژگی استفاده کرده و طرح‌های اینترنت حجمی را به این گروه ارائه نمایند. از بررسی شکل (۶) مشخص است که تقریباً ۸۵ درصد از کل کاربران از سرویس ایمیل استفاده می‌نمایند. می‌توان از این ویژگی برای ارسال تبلیغات در ایمیل مشتریان استفاده نمود. در ادامه با بررسی شکل (۷) می‌توان به این مسئله پی برد که افراد بسیار کمی از فایل‌های ویدئویی استفاده می‌نمایند بنابراین بهتر است از فایل‌های ویدئویی جهت تبلیغات استفاده نشود. با بررسی شکل (۸) می‌توان متوجه شد که تقریباً ۶۰ درصد افراد از وب برای اطلاعات پزشکی استفاده نمی‌نمایند این مسئله می‌تواند ناشی از نبود محتوای مناسب و عدم آگاهی افراد در این زمینه باشد.

٦- نتیجه‌گیری

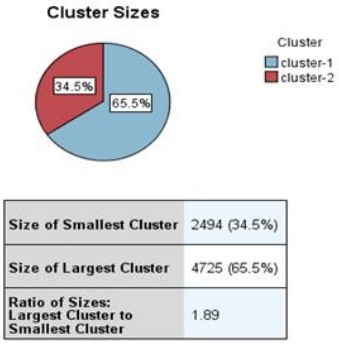
در این مقاله به بررسی الگوریتم‌های خوشه‌بندی k-میانگین، کوهونن و الگوریتم خوشه‌بندی دومرحله‌ای جهت شناسایی کاربرانی با ویژگی‌های مشابه پرداختیم. کیفیت خوشه‌بندی در هر سه روش پایین بوده و جهت افزایش کیفیت از روش تحلیل مؤلفه اصلی برای کاهش بعد استفاده نمودیم. با استفاده از این روش به معیار سیلهوت مطلوبی رسیده و کیفیت

خوشه‌بندی افزایش یافت. در بین این سه الگوریتم k–میانگین بهترین عملکرد و بالاترین میزان سیلهوت را دارا بود. با بررسی و تحلیل نتایج حاصل از هر خوشه می‌توان خدمات بهتر و سفارشی به کاربران ارائه نمود.

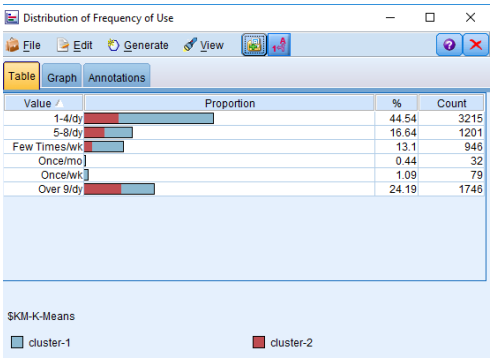
ضمایم



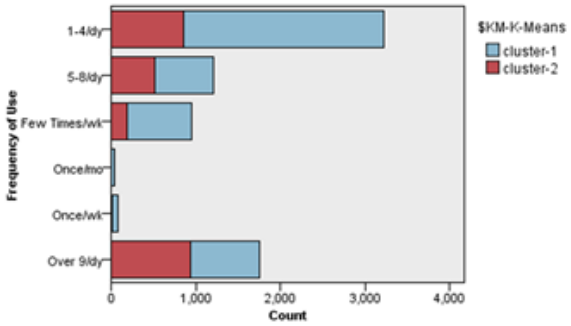
شکل (۲): ارزیابی کیفیت خوشه‌بندی



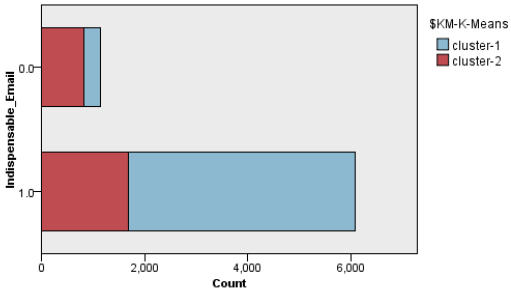
شکل (۳): خوشه‌بندی مجموعه داده به دو خوشه



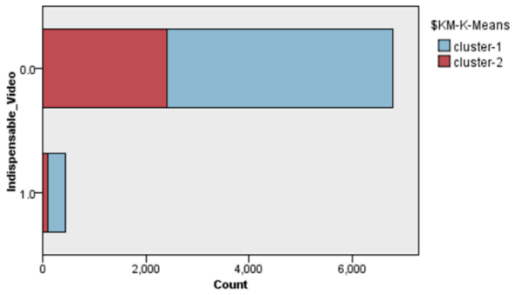
شکل (٤): تعداد دفعات استفاده از وب



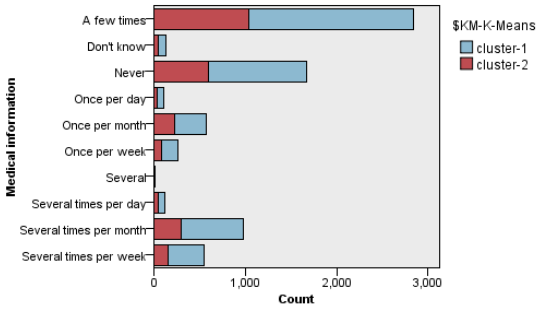
شکل (۵): تعداد دفعات استفاده از وب



شکل (٦): استفاده از ایمیل



شکل (۷): استفاده از ویدئو



شکل (۸): میزان استفاده از اطلاعات پزشکی بر روی وب

مراجع

[۱] غضنفری، مهدی؛ فتحیان، محمد؛ دشتی، یاسر **"استخراج الگوی حرکتی کاربران وبگاه با استفاده از تکنیک خوشه‌بندی و ارائه ساختار ماشین پیشنهاددهنده"**، پنجمین کنفرانس ملی مهندسی صنایع, ۱۳۹۶.

[۲] روحانی، سمیه؛ ملک‌زاده، مینا؛ بیات، مرتضی **"خوشه‌بندی نتایج جستجوی وب بر اساس علایق کاربران"**، همایش ملی سیستم‌های هوشمند در مهندسی برق و کامپیوتر, ۱۳۹۳.

[3] R. Kosala and H. Blockeel, "*Web mining research: A survey*," ACM Sigkdd Explorations Newsletter, vol. 2, pp. 1-15, 2000.

[4] B. Singh and H. K. Singh, "*Web data mining research: a survey*," in Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on, 2010, pp. 1-10.

[5] J. Zhang, P. Zhao, L. Shang, and L. Wang, "*Web usage mining based on fuzzy clustering in identifying target group*," in Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on, 2009, pp. 209-212.

[6] A. Abraham and V. Ramos, "*Web usage mining using artificial ant colony clustering and linear genetic*

programming," in Evolutionary Computation, 2003. CEC'03. The 2003 Congress on, 2003, pp. 1384-1391.

[7] M. Belk, E. Papatheocharous, P. Germanakos, and G. Samaras, "*Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques*," Journal of Systems and Software, vol. 86, pp. 2995-3012, 2013.

[8] Y. Fu, K. Sandhu, and M.-Y. Shih, "*Clustering of web users based on access patterns*," in Proceedings of the 1999 KDD Workshop on Web Mining, 1999.

[9] P. Lingras and C. West, "*Interval set clustering of web users with rough k-means*," Journal of Intelligent Information Systems, vol. 23, pp. 5-16, 2004.

[10] B. Mobasher, "*Data mining for web personalization*," in The adaptive web, ed: Springer, 2007, pp. 90-135.

[11] O. Etzioni, "*The World-Wide Web: quagmire or gold mine?*," Communications of the ACM, vol. 39, pp. 65-68, 1996.

[12] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "*User profiles for personalized information access*," in The adaptive web, ed: Springer, 2007, pp. 54-89.

[13] P. Berkhin, "*A survey of clustering data mining techniques*," in Grouping multidimensional data, ed: Springer, 2006, pp. 25-71.

[14] P. Prabhu and N. Anbazhagan, "*Improving the performance of k-means clustering for high dimensional data set*," International journal on computer science and engineering, vol. 3, pp. 2317-2322, 2011.

[15] L. Van Der Maaten, E. Postma, and J. Van den Herik, "*Dimensionality reduction: a comparative*," J Mach Learn Res, vol. 10, pp. 66-71, 2009.

[16] D. S. Sisodia, S. Verma, and O. P. Vyas, "*Augmented intuitive dissimilarity metric for clustering of web user sessions*," Journal of Information Science, vol. 43, pp. 480-491, 2017.

[17] X. Xie and B. Wang, "*Web page recommendation via twofold clustering: considering user behavior and topic relation*," Neural Computing and Applications, vol. 29, pp. 235-243, 2018.

[18] A. Morichetta, E. Bocchi, H. Metwalley, and M. Mellia, "*CLUE: Clustering for Mining Web URLs*," in Teletraffic Congress (ITC 28), 2016 28th International, 2016, pp. 286-294.

[19] T. Hussain and S. Asghar, "*Chi-square based hierarchical agglomerative clustering for web sessionization*," Journal of the National Science Foundation of Sri Lanka, vol. 44, 2016.

[20] Available: https://www.cc.gatech.edu/gvu/user_surveys/survey-1997-10/datasets/final_use.repl

[21] Available: https://www.cc.gatech.edu/gvu/user_surveys/survey-1997-10/questions/use.html

[22] H. Abdi and L. J. Williams, "*Principal component analysis*," Wiley interdisciplinary reviews: computational statistics, vol. 2, pp. 433-459, 2010.

[23] I. T. Jolliffe, "*Principal component analysis and factor analysis*," Principal component analysis, pp. 150-166, 2002.

[24] C. C. Aggarwal and C. K. Reddy, *Data clustering: algorithms and applications*: CRC Press, 2013.

[25] T. Kohonen and T. Honkela, "*Kohonen network*," Scholarpedia, vol. 2, p. 1568, 2007.

[26] I. S. Modeler, "*14.2 Algorithms Guide*," IBM Corporation, 2011.

[27] S. Bollmann, A. Hölzl, M. Heene, H. Küchenhoff, and M. Bühner, "*Evaluation of a new k-means approach for exploratory clustering of items*," 2015

زیر نویس

[1] Etzioni
[2] Chi-square
[3] Hierarchical Methods
[4] Logs
[5] Partitioning Methods
[6] Density-Based Methods
[7] TwoStep
[8] Kohonen
[9] Euclidean Distance
[10] Log-Likelihood
[11] Silhouette