دانشگاه آزاد اسلامی واحد اهر
فصلنامه‌ی علمی فضای جغرافیایی

سال بیست¬و¬سوم، شماره‌ی ۸۴
زمستان ۱۴۰۲، صفحات ۱۶۸-۱۴۵

*انسیه میهن پرست [۱]
جواد سدیدی [۲]

داده کاوی الگوی سفرمسافران با استفاده از اطلاعات داوطلبانه مکانی (مطالعه موردی: شهر تهران)

تاریخ دریافت: ۱۴۰۰/۰۲/۰۴ تاریخ پذیرش: ۱۴۰۰/۰۸/۲۶

چکیده

باتوجه به سفرهای صورت گرفته در شهر تهران و شناسایی مکان‌های پربازدید برحسب علایق گردشگران، پژوهش حاضر در راستای داده کاوی الگوی رفتاری سفر مسافران می باشد. در این جهت هدف پژوهش حاضر، شناسایی مکان‌های پربازدید مسافران شهرتهران از روی رفتار آن ها برای مقاصد گردشگری است تا به کمک این الگوها بتوان الگوی سفر آن ها را بر اساس اطلاعات داوطلبانه جغرافیایی بررسی و ارزیابی کرد. بدین منظور از داده کاوی و مدل RFMبرای تجزیه و تحلیل روابط بین مشخصه‌های مسافران و تمایلات آن ها استفاده شده است. جهت انجام مراحل عملیاتی پژوهش نیاز به اپلیکیشن قابل نصب بر تلفن همراه و سرور ذخیره سازی داده‌های ارسالی از سمت گردشگر بوده که توسط پژوهشگر آماده سازی گردید و در گام بعدی در دسترس گردشگران شهر تهران قرار گرفته است. پس از ذخیره سازی داده‌های داوطلبانه جغرافیایی توسط گردشگر براساس آخرین آمارهای گردشگری نزدیک ۸۵٪ مکان‌های پربازدید استان تهران، حسب سلیقه و بازدیدهای گردشگران ثبت گردیده که شامل سفرهای مذهبی، تفریحی، فرهنگی- تاریخی، علمی و اجتماعی می شود. باتوجه به نحوه جمع آوری داده ها، در مدل RFM اقدام به بررسی انواع سفرها، تعداد مکان‌های مورد بازدید و اطلاعات گردشگران شده است و با شناسایی ساختار داده‌ای موجود، اقدام به تطبیق داده‌های دریافتی با متغیرهای تاریخ بازدید از مکان (جدیدترین تاریخ بازدید از مکان) R و (مدت زمان بازدید از مکان) M و تعداد دفعات بازدید از مکان (تکراربازدید از مکان) F گردیده است. روش RFM امکان خوشه بندی و تحلیل داده‌ها را جهت دستیابی به الگوی سفر گردشگران در اختیار محقق قرار می دهد و مدل مناسبی جهت تحلیل اطلاعات می باشد. نتایج پژوهش حاکی از آن است که براساس علایق گردشگران شهر تهران، مکان‌های تفریحی در راس هرم قرار دارند که براساسCLV محاسبه شده در پژوهش پلاتینی نامیده می شوند و مکان‌های فرهنگی و تاریخی و مذهبی و زیارتی در درجه دوم نقره‌ای می باشد و مکان‌های اجتماعی در درجه سوم مکان‌های مسی نامیده می‌شوند که این طیف درجه بندی

۱- دانشجوی کارشناسی ارشد، سنجش از دور و سیستم اطلاعات جغرافیایی ، دانشگاه خوارزمی، تهران ،ایران،(نویسنده مسئول)
E-mail: yazdani.m51@gmail.com
*۲- گروه سنجش از دور و سیستم اطلاعات جغرافیایی،، دانشکده علوم جغرافیایی، دانشگاه خوارزمی، تهران، ایران،
E-mail: jsadidi@gmail.com

نشان از اهمیت مکان‌های بازدیدشده برای گردشگران دارد. لازم به ذکر است که برای معرفی مکان‌های پر بازدید به گردشگران دیگر نیز می توان از سامانه گردشگری استان تهران جهت جذب گردشگر به این نوع مکان ها بهره برد.

کلمات کلیدی: اطلاعات داوطلبانه جغرافیایی – الگوی سفر – داده کاوی – توریسم – روش RFM

مقدمه

بررسی مکان‌های پربازدید، از جمله مهم ترین موضوعات موجود بر سر راه داده‌های جغرافیایی داوطلبانه در راستای بررسی الگوی سفرگردشگران می‌باشد. شناسایی الگوی سفر (تشخیص الگو) شاخه‌ای از مبحث یادگیری ماشین است. می‌توان گفت تشخیص الگو، دریافت داده‌های خام و تصمیم‌گیری بر اساس دسته‌بندی داده‌ها است. تشخیص الگو می‌تواند به عنوان دسته‌بندی داده‌های ورودی در کلاس‌های شناخته شده به وسیله استخراج ویژگی‌های مهم تعریف شود. بر این اساس مردم اطلاعات خود را که با هدف مشخصی جمع‌آوری شده اند را در اختیار سیستم قرار می دهند. در این نگرش هر فرد می تواند مانند سنجنده‌ای باشد که محیط پیرامون خود را رصد میکند. از جمله انگیزه‌های اصلی برای رفتن به سمت چنین فرآیندهایی، دسترسی به حجم بالای اطلاعات و هزینه پایین دستیابی به منابع دقیق اطلاعات جغرافیایی می باشد. بنابراین، هدف تحقیق ارائه مدلی مفهومی است که با استفاده از آن بتوان داده‌های ورودی به پایگاه داده داوطلبانه را به صورت درست ذخیره سازی کرد و براساس اطلاعات داوطلبانه گردشگران شهر تهران الگوی سفر آنان را مورد ارزیابی قرار داد.

Pashaei and Malik ,(2014) در پژوهش خود تحت عنوان بررسی کیفیت اطلاعات مکانی داوطلبانه از منظر شاخص با تاکید بر شاخص برازندگی، به این نتیجه رسیده اند که اطلاعات مکانی داوطلبانه گونه ای از اطلاعات است که به صورت اختیاری توسط مردم عادی، بدون نیاز به آموزش علمی و با توجه به دانش محلی تهیه می شود. داده‌های داوطلبانه مکانی در مقایسه با داده‌های مکانی استاندارد، دارای مزایای فراوانی از جمله دسترسی آسان اطلاعات و به روزرسانی سریع است. ولی کیفیت این داده‌ها یکی از بحث‌های مهم در این راستاست.

Pashaei& Malek 1392)) در مطالعه موضوع محیط‌های اطلاعات مکانی مردم گستر(ویژگی ها و چالش ها) به این نتیجه رسیده اند که اهمیت داده‌های داوطلبانه مکانی مبتنی بر این حقیقت است که مردم مخصوصا ساکنین یک محل بیشترین شناخت نسبت به توصیف زندگی خود را دارند و عموما این افراد بدون دانش نقشه‌برداری هستند. که بر خلاف داده‌های مکانی استاندارد که از فراداده استفاده می کنند.

پیشینه پژوهش

(Hansen,2011) در مورد مطالعاتی خود با عنوان مدیریت اطلاعات داوطلبانه مکانی – مطالعه موردی در مورد اجرای ۳SDI به این نتیجه رسیده است که: از اهداف تولید و استفاده از منابع داده های داوطلبانه جغرافیایی ؛ تقویت، بهنگام سازی و تکمیل پایگاه‌های داده مکانی موجود است. دسترسی و به اشتراک گذاری داده‌های مکانی، همیشه به عنوان یک چالش بزرگ برای محققین مطرح بوده است.

Farzanyar and Cercone,(2016) در مقاله خود با موضوع استخراج الگوی سفر با استفاده از عکس‌های دارای برچسب جغرافیایی در مقیاس بزرگ ، دریافتند که وب سایت‌های اشتراک گذاری عکس به افراد این امکان را می دهد تا تجربیات خود را از طریق وب از طریق داده‌های غنی رسانه‌ای مانند عکس به وب نمایش دهند. این عکس‌ها از نظر عرض و طول جغرافیایی که در آن عکس گرفته شده، دارای فضای مکانی هستند. عکس‌های جغرافیایی دارای اطلاعات زیادی درباره رفتار مسافرتی افراد و تراکم مکان‌های گردشگری هستند. مانند فناوری‌های مبتنی بر وب و مبتنی بر تلفن همراه، عکس‌های دارای برچسب جغرافیایی به طور فزاینده‌ای فراتر از توانایی تحلیل انسانی جمع می شوند. بنابراین، فرصت ها و چالش‌های تحقیقاتی جدیدی را ارائه داده‌اید. تجزیه و تحلیل رفتارهای مسافرتی بر اساس عکس‌های دارای برچسب جغرافیایی٤ بررسی می شود.

Azizkhani and Malek (2016) در پژوهش خودکه دقت اطلاعات جغرافیایی داوطلبانه در زمینه لرزه‌های‌هایتی بوده است به این نتیجه رسیده اند که استفاده بهینه از سیستم اطلاعات مکانی در زمینه مدیریت داده‌ها، مستلزم داشتن داده‌های مناسب است. به دست آوردن و جمع‌آوری داده‌ها به طرق مختلفی قابل انجام است که عمده ترین و مرسوم ترین روش استخراج این است که هر سازمان یا ارگانی اقدام به جمع‌آوری این داده‌ها نموده و از آنها در مواقع مورد نیاز استفاده نماید. با توجه به وسعت کشورها، چنانچه سازمان مشخصی موظف به جمع‌آوری این اطلاعات شود، نیاز به صرف هزینه های هنگفت خواهد بود.

Goodchild (2011) در بررسی موضوع جمع‌آوری و داده کاوی اطلاعات جغرافیایی برای کنترل سفرهای غیر ضرور شهری دریافتند که از انگیزه‌های اصلی برای رفتن به سمت داده کاوی و اطلاعات داوطلبانه جغرافیایی، عدم دسترسی و هزینه بالای دستیابی به منابع دقیق اطلاعات جغرافیایی از روش های مرسوم می‌باشد. با استفاده از اطلاعات داوطلبانه انبوهی از اطلاعات و نقشه ها در مدت زمان کوتاهی تهیه می گردند. بدین ترتیب سازمان ها و ارگان ها می توانند تنها بر روی کمبودها و نیازهای خود که به شکل توزیع یافته و داوطلبانه قابل دستیابی نیست، تمرکز نمایند. علاوه برآن بستر اینترنت و برنامه های کاربردی مبتنی بر شبکه در حال حاضر توسط وسایل همراه کاربران و سیستم های عامل پرطرفداری مانند Android در دسترس هستند.

3 Spatial Data Infrastructure
4 Photos with Geotags

Cloner (2011) در مقاله خود در تجزیه و تحلیل مسیر الگوی سفرهای درون شهری (مرور سیستماتیک رویکردهای فعلی شهری) به این نتیجه رسیده اند که سه چالش مهم که سودمندی داده‌های داوطلبانه جغرافیایی را تحت تأثیر قرار می‌دهند را شناسایی کرده اند. این چالش ها عبارتند از جمع‌آوری داده‌ها، تعیین موقعیت و ارزیابی کیفیت آنها. آن ها اذعان نموده اند که تا عدم رفع این چالش ها نمی توان سودمندی بالایی از داده‌های داوطلبانه حاصل نمود.

Mozafari(1396) در مقاله بخش بندی مشتریان بانک بر اساس ارزش دوره عمر آن ها و مدل RFM با ترکیب روش‌های تصمیم گیری چندمعیاره و داده کاوی دریافته است که ابتدا داده‌های مربوط به شاخص‌های مدل RFM از پایگاه داده بانک استخراج و پیش پردازش می‌شود. پس از آن وزن هر کدام از شاخص ها با استفاده از روش آنتروپی شانون۵ محاسبه میگردد. سپس مشتریان به کمک شبکه عصبی خودسازمان ده٦ به پنج بخش یا خوشه اصلی تفکیک شده و ویژگی‌های مشتریان در هر یک از بخش ها شناسایی میشوند. در نهایت هرم ارزش دوره عمر مشتری تشکیل می شود.

منطقه مورد مطالعه

تهران پرجمعیت‌ترین شهر و پایتخت ایران، مرکز استان تهران و شهرستان تهران است. با ۸،٦۹۳،۷۰٦ تن جمعیت، بیست و چهارمین شهر پرجمعیت جهان و پرجمعیت‌ترین شهر باختر آسیا به‌شمار می‌رود .کلان‌شهر تهران نیزدومین کلان‌شهر پرجمعیت خاورمیانه است. تهران کانون اقتصادی ایران است و اولین منطقه صنعتی این کشور محسوب می‌شود اما فعالیت‌های اقتصادی بین‌المللی نقش چندانی در شمار شاغل‌های آن ندارد. این شهر یکی از مهم‌ترین مراکز گردشگری ایران به حساب می‌آید و دارای مجموعه‌ای از جاذبه‌های گردشگری است که شامل کاخ‌ها و موزه‌هایش می‌شود(National Statistics Organization, 2018).

شکل ۱ : موقعیت استان تهران و شهرستان های اطراف

Figure 1 : Geographical location of Tehran province and suburbs

5 Shannon Entropy
6 Self-Organizing Neural Network

تهران در پهنه‌ای بین دو وادی کوه و کویر و در دامنه‌های جنوبی البرز گسترده شده‌است و ۷۳۰ کیلومتر مربع مساحت دارد. از نظر جغرافیایی نیز در ۵۱ درجه و ۱۷ دقیقه تا ۵۱ درجه و ۳۳ دقیقه طول خاوری و ۳۵ درجه و ۳۶ دقیقه تا ۳۵ درجه و ٤٤ دقیقه عرض شمالی قرار دارد شکل (۱). همچنین بافت نامتراکم، وجود باغ‌های کهن، بوستان‌ها، فضای سبز حاشیه بزرگراه‌ها و کم بودن فعالیت‌های صنعتی در شمال شهر کمک کرده‌اند تا هوای مناطق شمالی به‌طور متوسط ۲ تا ۳ درجه سانتی‌گراد خنک‌تر از مناطق جنوبی شهر باشد. استان تهران با جذب ٦٤/١ میلیون گردشگر در سال ۲۰۱٦ میلادی، یکی از مهم‌ترین شهرهای خاورمیانه در زمینه گردشگری بوده‌است. همچنین تهران پس از شهرهای دبی، ژوهانسبورگ، ریاض و ابوظبی، در رتبه پنجم از دیدگاه شمار ورود گردشگران خارجی در سال ۲۰۱٦ در مناطق آفریقا و خاورمیانه قرار داشت و گردشگران خارجی تهران در این سال، نیم میلیارد دلار هزینه‌کرد داشته‌اند(National Statistics Organization, 2018).

مواد و روش پژوهش

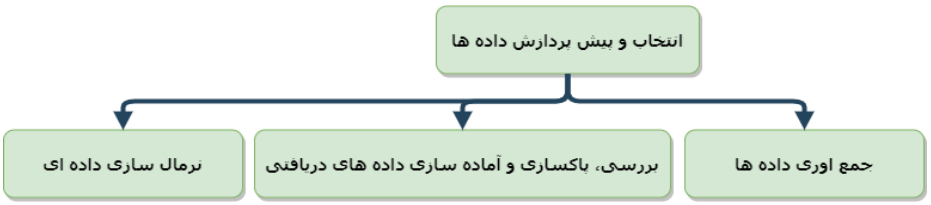
الف) مواد و داده های پژوهش

تجزیه و تحلیل داده‌ها (داده کاوی) فرآیندی چندمرحله‌ای‌است که طی آن داده‌هایی که از طریق بکارگیری ابزارهای جمع‌آوری در نمونهٔ (جامعه) آماری فراهم‌آمده‌اند؛ خلاصه، کدبندی و دسته‌بندی و در نهایت پردازش می‌شوند تا زمینهٔ برقراری انواع تحلیل‌ها و ارتباط‌ها بین این داده‌ها به‌منظور آزمون فرضیه‌ها فراهم آید. در این فرآیند داده‌ها هم از لحاظ مفهومی و هم از جنبهٔ تجربی پالایش می‌شوند و تکنیک‌های گوناگون آماری نقش بسزایی در استنتاج‌ها و تعمیم‌ها به عهده دارند. Khaki,(1387) در این بخش داده های جمع آوری شده مورد ارزیابی قرار گرفته است. گسترش تکنولوژی و ظهور امکانات جدید در حوزه اینترنت، بستری را برای تولید داده‌های مکانی توسط عموم مردم و بصورت داوطلبانه فراهم کرده است. بدین ترتیب داده‌های مکانی بر خلاف روند سنتی تولید داده، توسط هر کاربر تولید شده و بصورت رایگان در دسترس سایرین قرار می‌گیرد. این پدیده با نام اطلاعات مکانی این پدیده تأثیر بسزایی در تولید و به اشتراک‌گذاری داده‌های مکانی داشته، بطوریکه هر فرد می‌تواند هم تولید کننده و هم کاربر داده‌های مکانی باشد. VGI منبعی غنی و ارزشمند از داده‌های مکانی بوده که به کاربران اجازه می‌دهد تا دنیا را بر اساس درک و چشم‌انداز خود تصویر کنند. داده‌های پژوهش حاضراز پایگاه داده اطلاعاتی SQL که بااستفاده از اطلاعات داوطلبانه کاربران ذخیره سازی شده است استخراج گردیده، به منظور انجام این پژوهش حدود ۳۰۰ نمونه از این پایگاه استخراج گردیده و در اختیار محقق قرار گرفته است.

شکل ۲: اپلیکیشین قابل نصب بر روی تلفن همراه جهت جمع‌آوری اطلاعات سفر مسافران

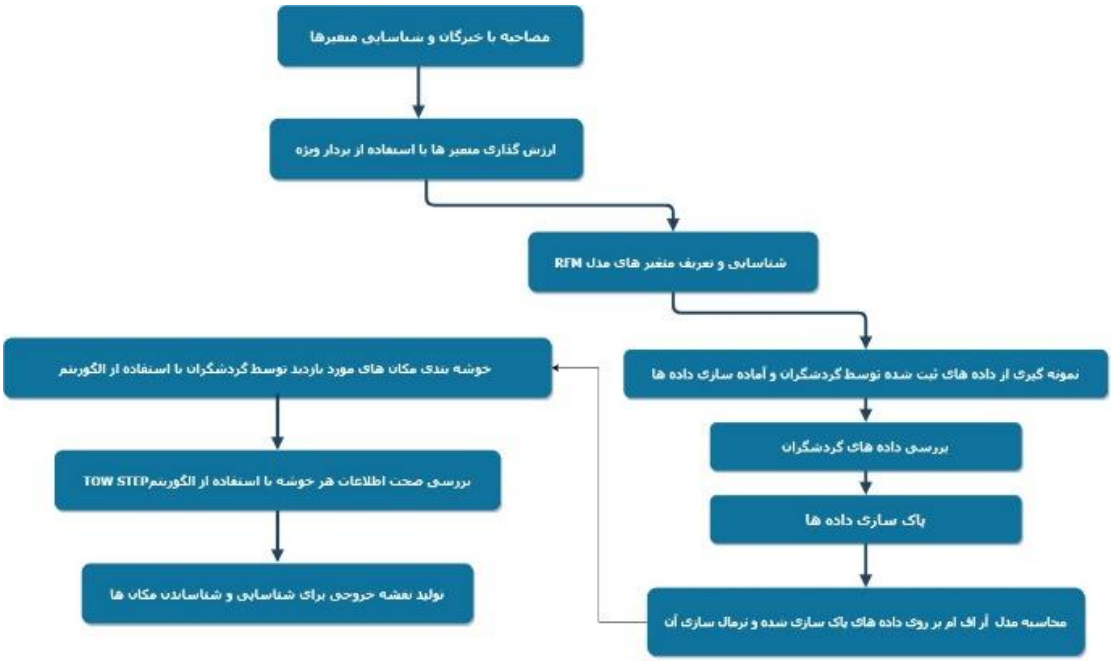
Figure 2: An application that can be installed on a mobile phone to collect travel information of passengers

ب) روش پژوهش



شکل ۳: دیاگرام انتخاب و پردازش داده‌ها

Figure 3: Data selection and processing diagram



شکل ٤: ساختار کلی مدل اجرایی پژوهش

Figure 4: The overall structure of the research executive model

همانطور که در شکل (٤)دیده می شود فرآیند اصلی برای خوشه بندی مکان های مورد بازدید گردشگران در این مدل شامل مراحل اصلی زیر می باشد:

۱- محاسبه ضریب اهمیت شاخص های RFM

۲- انتخاب متغیر های مناسب و استخراج نمونه

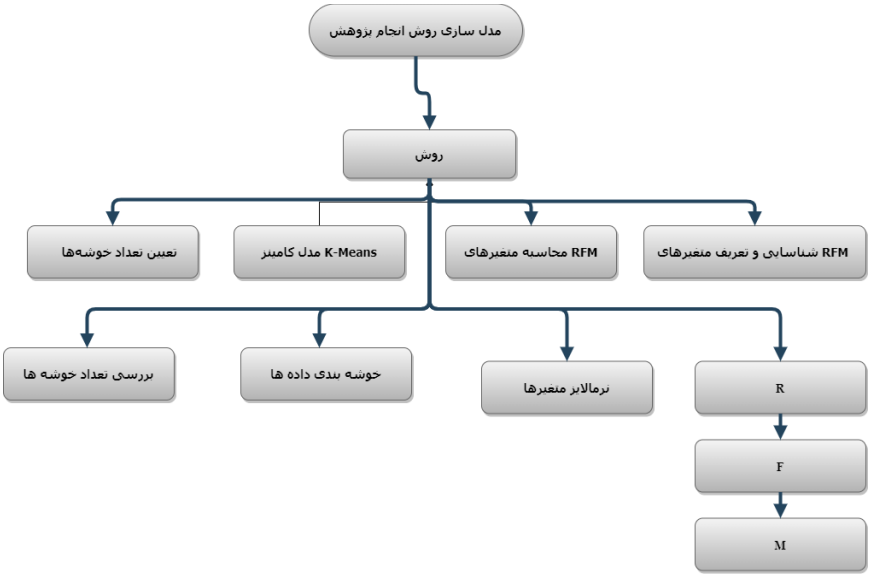
۳- آماده سازی و پاک سازی داده ها داده ها

٤- ساختار مفهومی خوشه بندی مکان های مورد بازدیدگردشگر و متغیر های RFM

۵- ارزیابی مدل با استفاده از الگوریتم TowStep وارزیابی شخص clv

٦- دسته بندی مناطق مورد بازدید گردشگران با استفاده از داده کاوی و ترسیم نقشه مکان های موردبازدید.

در این پژوهش با توجه به حجم بالای داده‌ها و تنوع و به عبارتی تعدد فیلدهای اطلاعاتی در قبال هر مسافر از نرم افزار SQL Server 2015 به منظور ذخیره سازی داده‌های استفاده گردیده است. داده‌های لازم در این پژوهش با استفاده از اطلاعات داوطلبانه جغرافیایی توسط هر شخص (کاربر) اخذ گردیده است. داده‌های خام موجود در پایگاه داده‌ای اغلب به شکل پردازش نشده، ناقص و نویزی هستند. برخی محدودیت‌های موجود در پایگاه داده‌ای مواردی نظیر وجود داده‌های قدیمی۷ یا زاید۸، مقادیر مفقود، داده‌های پرت۹، شکل نامناسب داده‌ها برای داده کاوی هستند. از این رو برای آماده سازی این داده‌ها برای کاربرد در اهداف داده کاوی، نیاز به اجرای گام‌هایی نظیر پاکسازی داده‌ای۱۰ و تبدیل داده‌ای۱۱ می باشد.



شکل ۵ : دیاگرام مدل سازی روش پژوهش

Figure 5: Research method modeling diagram

7 Obsolete
8 Redundant
9 Outlier
10 Data cleaning
11 Data transformation

یکی از مدل‌های ساده و در عین حال قدرتمند در پیاده سازی داده کاوی الگوی سفر که کاربرد فراوانی در صنعت گردشگری دارد، مدل RFM می باشد. در این مرحله اقدام به بررسی انواع سفرها، تعداد مکان‌های مورد بازدید و اطلاعات گردشگران می‌گردد، و با شناسایی ساختار داده‌ای موجود در اقدام به تطبیق داده‌های دریافتی با متغیرهای تاریخ بازدید از مکان (جدیدترین تاریخ بازدید از مکان) R و مدت زمان بازدید از مکان M و تعداد دفعات بازدید از مکان (تکرار بازدید از مکان) F می گردد. با توجه به داده‌های پاکسازی شده اقدام به پیاده سازی دستورات لازم به منظور محاسبه متغیرهای R و F و M برای مکان ها با استفاده از داده‌های داوطلبانه گردشگران می گردد. خروجی این مرحله داده‌های محاسبه شده و نرمال شده، برای انتقال به نرم افزار مربوطه جهت عملیات خوشه بندی سفرها می باشد. در رابطه با نرمالایز کردن تاریخ در RFM واژه‌ی Recently به معنای تازگی (تازگی مکان) وجود دارد. زمانی که داده‌های داوطلبانه از سمت کاربر جمع‌آوری می شود هریک از این داده‌ها تاریخی دارند با عنوان تاریخ ثبت اطلاعات، یعنی یک کاربر داده‌ای را در ٦ ماه قبل در یک تاریخ ثبت و یک کاربر داده‌ای را در ۲ ماه قبل در تاریخی دیگر ثبت نموده است. اینکه تاریخ زیادی از ثبت اطلاعات مکانی گذشته باشد به عنوان داده‌های پرت شناخته می شوند. حال برای شناسایی داده‌های پرت، تاریخ ها را باید به عدد تبدیل کرد، یعنی به عبارتی تاریخ ها را به داده‌های کمی تبدیل کرد. حال سوال پیش می آید که چگونه؟ حال در این بخش آخرین بازدید از هر مکان ملاک قرار میگیرد و تاریخ هریک از اطلاعات ثبت شده براساس آخرین تاریخ سنجیده می‌شود. زمان معرفی داده‌ها به نرم افزار SPSS Celementine داده‌ها را تبدیل می‌شوند به عبارتی نرمالایز می گردند و تاریخ هر بازدید را مورد محاسبه قرار میگیرد. پارامتر تاریخ به عنوان Recently و یا تازگی بازدید از مکان از ارکان اصلی RFM است و به این علت از آن در پژوهش حاضر استفاده شده است. در نتیجه مقادیری که پس از این تبدیل تاریخ ها خارج از محدوده ۳- تا ۳+ قرار داشته باشند به عنوان داده‌های پرت حذف می شوند.

بر اساس موضوع پژوهش و نظر خبرگان تمرکز این داده کاوی الگوی سفر مسافران با استفاده از اطلاعات داوطلبانه جغرافیایی، مطالعه موردی: شهر تهران بوده و هدف خوشه بندی مکان‌های مورد بازدید گردشگر در استان تهران می باشد، بدین منظور متغیرهای زیر مورد نظر قرار گرفت:

۱- مشخصه تازگی مکان ها R : فاصله زمانی بین آخرین مکان ثبت شده توسط گردشگر تا پایان دوره ثبت اطلاعات در اپلیکیشن.

۲- تکرار مکان‌های مورد بازدید F : تعداد مکان‌هایی است که یک گردشگر در یک دورهٔ زمانی خاص سفرکرده است.

۳- مدت زمان بازدید از مکان ها M : مدت زمان بازدیدهای صورت گرفته توسط گردشگر.

همچنین خوشه بندی براساس مدل K-Means که یکی از روش‌های خوشه بندی داده‌ها در داده کاوی است، صورت گرفته است. این روش علی‌رغم سادگی آن یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر (مانند خوشه‌بندی فازی) محسوب می‌شود. بدست آوردن نقاطی به عنوان مراکز خوشه‌ها این نقاط در واقع همان میانگین نقاط متعلق به هر خوشه هستند. نسبت دادن هر نمونه داده به یک خوشه که آن داده کمترین فاصله تا مرکز آن خوشه را دارا باشد. در نوع ساده‌ای از این روش ابتدا به تعداد خوشه‌های مورد نیاز نقاطی به صورت تصادفی انتخاب می‌شود. سپس در داده‌ها با توجه با میزان نزدیکی (شباهت) به یکی از این خوشه‌ها نسبت داده می‌شوند و بدین ترتیب خوشه‌های جدیدی حاصل می‌شود. با تکرار همین روال می‌توان در هر تکرار با میانگین‌گیری از داده‌ها مراکز جدیدی برای آنها محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. این روند تا زمانی ادامه پیدا می‌کند که دیگر تغییری در داده‌ها حاصل نشود. تابع زیر به عنوان تابع هدف مطرح است. در الگوریتم K-Means ابتدا k عضو (که k تعداد خوشه‌ها است) بصورت تصادفی از میان n عضو به عنوان مراکز خوشه‌ها انتخاب می‌شود. سپس n-k عضو باقیمانده به نزدیک‌ترین خوشه تخصیص می‌یابند. بعد از تخصیص همه اعضا مراکز خوشه مجدداً محاسبه می‌شوند و با توجه به مراکز جدید به خوشه‌ها تخصیص می‌یابند و این کار تا زمانی که مراکز خوشه‌ها ثابت بماند ادامه می‌یابد (Farzanyar and cercone) .

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

معادله ۱ : معادله تکنیک K-Means

Formula1: Equation of K-Means technique

جدول ۱ – اطلاعات ثبت شده در اپلیکیشن

Table 1: Information recorded in the application

تعداد اطلاعات ثبت شده در اپلیکیشن	تعداد اطلاعات صحیح ثبت شده	تعداد اطلاعات نادرست ثبت شده
۳۰٤	۲۹٤	۱۰

جدول ۱:توزیع فراوانی پاسخ دهندگان برحسب جنسیت

Table 2: Frequency distribution of respondents according to gender

جنسیت	فراوانی	درصد فراوانی	درصد فراوانی تجمعی
زن	۷۷	٪۵۱	٪۵۱
مرد	۷۵	٪٤۹	٪۱۰۰
جمع کل	۱۵۲	٪۱۰۰	---

جدول۳ : توزیع فراوانی داده ها برحسب موقعیت شهر تهران و اطراف آن

Table 3: Frequency distribution of data according to the location of Tehran and its surroundings

موقعیت شهر	فراوانی	درصد فراوانی	درصد فراوانی تجمعی
تهران	۸۵	٪۸۲	٪۸۲
شهرستان های اطراف	۱۹	٪۱۸	٪۱۰۰
جمع کل	۱۰٤	٪۱۰۰	---

جدول ٤: توزیع فراوانی برحسب نوع گردشگری پاسخ دهندگان

Table 4: Frequency distribution according to the type of tourism of the respondents

انواع گردشگری	فراوانی	درصد فراوانی	درصد فراوانی تجمعی
مذهبی وزیارتی	۸	٪۸	٪۸
تفریحی	۳۲	٪۳۱	٪۳۸
فرهنگی و تاریخی	۲۱	٪۲۰	٪۵۹
علمی	۵	٪۵	٪٦۳
اجتماعی	۲۱	٪۲۰	٪٤
اقتصادی	۱۷	٪۱٦	٪۱۰۰

جدول ۵ : توزیع فراوانی سن پاسخ دهندگان

Table 5: Frequency distribution of age of respondents

سن	فراوانی	درصد فراوانی	درصد فراوانی تجمعی
کمتر از۲۵ سال	۲۹	٪۱۹	٪۱۹
بین ۲۵تا ۳۵سال	۸۵	٪۵٦	٪۷۵
بین ۳۵تا ٤۵ سال	۳۷	٪۲٤	٪۹۹
بیشتر از٤۵	۱	٪۱	٪۱۰۰
جمع کل	۱۵۲	٪۱۰۰	-----

جدول ٦: مقایسات زوجی مکان‌های مورد بازدید توسط گردشگر و متغیرهای مدلRFM

Table 6: Pairwise comparisons of places visited by tourists and RFM model variables

انواع گردشگری	مذهبی و زیارتی	تفریحی	فرهنگی و تاریخی	علمی	اجتماعی	اقتصادی
مذهبی وزیارتی	۱	۳	۵	۵	۳	۳
تفریحی	۱/۳	۱	۷	۸	۶	۷
فرهنگی و تاریخی	۱/۵	۱/۷	۱	۵	۳	۳
علمی	۱/۵	۱/۸	۱/۵	۱	۳	۳
اجتماعی	۱/۳	۱/۶	۱/۳	۱/۳	۱	۵
اقتصادی	۱/۳	۱/۷	۱/۳	۱/۳	۱/۵	۱

جدول ۷: ماتریس زوجی متغیرهای RFM

Table 7: Pair matrix of RFM variables

	تازگی R	تناوب F	حجم M
تازگی R	۱	۱/۲	۳
تناوب F	۲	۱	۳
حجم M	۱/۳	۱/۳	۱

به منظور محاسبه لاندای بیشینه می توان از روش نرم افزاری و یا محاسبات ریاضی استفاده نمود و در این پژوهش از روش نرم افزاری استفاده شده است، به منظور محاسبه لاندا در روش بردار ویژه از نرم افزار آنلاین Matrix Calculator و همچنین برای بدست آوردن وزن هر مشخصه با حل دستگاه معادلات از نرم افزار آنلاین Bahesab استفاده گردیده است. به منظور محاسبه وزن مکان‌های مورد بازدید بر اساس نظر خبرگان با استفاده از روش بردار ویژه ابتدا اقدام به محاسبه مقدار ویژه ماتریس مقایسات زوجی تشکیل شده با استفاده از Matrix Calculator گردیده است. مطابق با مقدار بردار ویژه بیشینه برای ماتریس مقایسات زوجی، درجه اهمیت مکان‌های مورد بازدید برابر است با :

$$\lambda_{\max} = 7.42$$

پس از محاسبه لاندای بیشینه اقدام به تشکیل دستگاه معادلات بر اساس فرمول زیر گردید:

$$(A - \lambda_{\max} I) \times W = 0$$

معادله ۲ :معادله جهت تشکیل دستگاه معادلات و محاسبه لاندا هرگروه

Formula 2: The equation to form the equation machine and calculate the Landa of each group

جدول ۸: لاندای بیشینه هر گروه

Table 8: The maximum landa of each group

X6	X5	X4	X3	X2	X1
۰/۰۳۸۰۶۶	۰/۰۰۶۶۶۱۶	۰/۰۷۰۴۲۳	۰/۱۲۲۵۷۳	۰/۳۴۳۷۳۸	۰/۳۵۸۵۸۴

مطابق با تصویر فوق وزن نوع مکان‌های مورد بازدید محاسبه گردید:

جدول ۹: وزن مکان‌های مورد بازدید

Table 9: Weight of visited places

اقتصادی	اجتماعی	علمی	فرهنگی و تاریخی	تفریحی	مذهبی و زیارتی
۰/۰۳۸۰۶۶	۰/۰۶۶۶۱۶	۰/۰۷۰۴۲۳	۰/۱۲۲۵۷۳	۰/۳۴۳۷۳۸	۰/۳۵۸۵۸۴

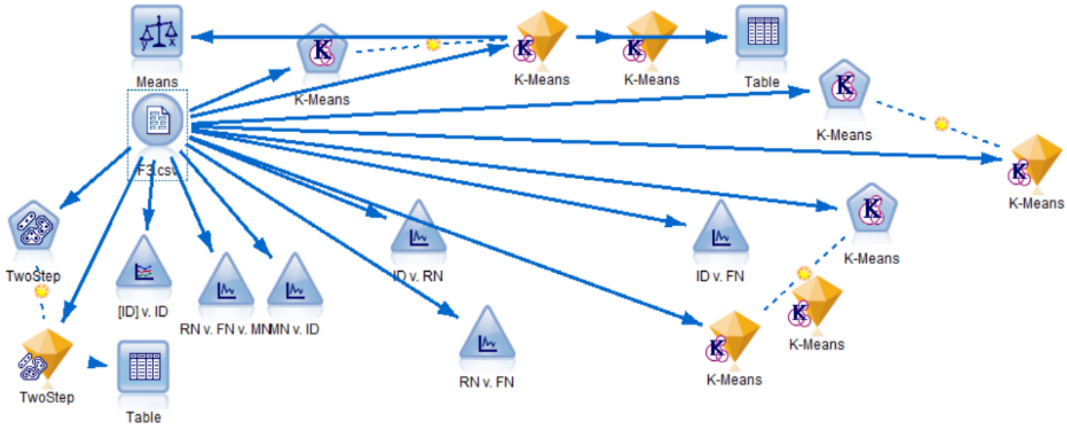
مطابق با وزن‌های محاسبه شده بیشترین ارجحیت مذهبی و زیارتی می باشد.

با فرآیند فوق و با روش بردار ویژه اقدام به محاسبه وزن متغیرهای مدل RFM گردیده و بر اساس نظر خبرگان وزن متغیر ها مطابق با جدول (۱۰) می باشد:

جدول ۱۰: وزن متغیرهای RFM

Table 10: Weight of RFM variables

تازگی R	تناوب F	حجم M
۰/۰۹	۰/۳۶	۰/۵۵



شکل ٦: مدل پیاده سازی شده پژوهش

Figure 6 : The implemented research model

٤ خوشه در جدول (۱۱) استخراج گردید و با اعمال ضرایب وزن‌های متغیرهای R و F و M اقدام به محاسبه درجه اهمیت بازدیدهای مکانی[12] برای هر خوشه می گردد در این جدول مرکز خوشه ها بر اساس متغیرهای R و F و M مشخص شده و CLVe مرکز هر خوشه نیز محاسبه گردیده است:

معادله محاسبه CLV_e: CLV=WR+WF+WM

جدول ۱۱ – جدول آنالیز خوشه ها

Table 11: Cluster analysis table

تعداد	عنوان	نرمال M	نرمال F	نرمال R	CLVe مرکز خوشه
۱۰۳	وزن	۰.۳۶	۰.۵۵	۰.۰۹	-
۲۷	خوشه اول	ضعیف	ضعیف	قوی	۰.۴۷۹۱۶۵۰۴۹
۱۹	خوشه دوم	ضعیف	ضعیف	قوی	۰.۰۶۷۱۶۸۲۷۷
٤٤	خوشه سوم	ضعیف	ضعیف	قوی	۰.۲۱۳۵۸۲۲۹۶
۱۳	خوشه چهارم	متوسط	متوسط	ضعیف	۰.۲۷۲۱۴۹۶۳۷

12 CLVe

جدول ۱۲- محاسبات جهت استخراج CLV

Table 12: Calculations to extract CLV

R-Rmin	F-Fmin	M-Mmin	Sum of RN	Sum of FN	Sum of MN	Cluster
۰	۰.۵۸۱۷۶۷۳۷	۰.۲۲۳۸	۰.۰۴۲۷	۰.۵۸۱۸	۰.۳۵۶۶	Cluster-1
۰.۰۳۰۳۴۸۰۲۵	۰.۱۸۲۹۰۴۵	۰	۰.۰۷۳۱	۰.۱۸۲۹	۰.۱۳۲۹	Cluster-2
۰.۲۳۵۸۸۱۰۰۴	۰.۲۳۵۳۲۸۱۳	۰.۰۸۲۵۱۷۴۸۳	۰.۲۷۸۶	۰.۲۳۵۳	۰.۲۱۵۴	Cluster-3
۰.۵۶۲۸۸۴۲۳۵	۰	۰.۱۶۲۲۳۷۷۶۲	۰.۶۰۵۶۰۵۹۱۹	۰	۰.۲۹۵۱	Cluster-4

Mmax-Mmin	=	۰.۲۲۳۷۷۶۲۲۴
Fmax-Fmin	=	۰.۵۸۱۷۶۷۳۷
Rmax-Rmin	=	۰.۵۹۲۲۹۱۰۶۸

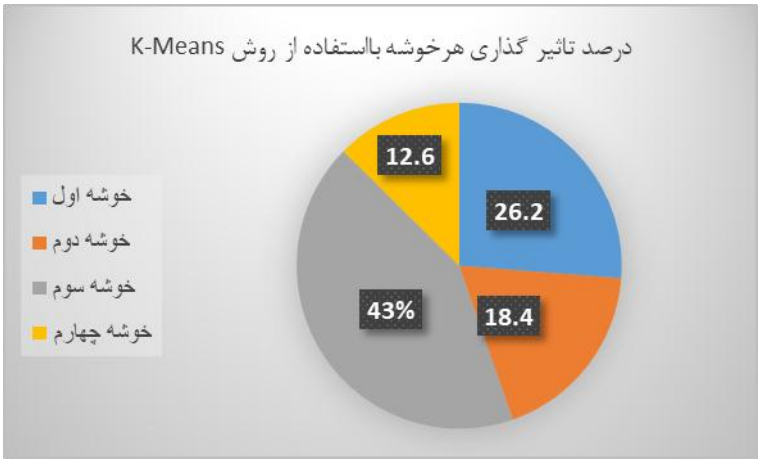
R''	F''	M''
۰.۰۷۴۵۳۹۸۰۶	۰.۲۰۱۴۹۵۱۴۶	۰.۲۷۷۶۶۹۹۰۳

CLV=WR+WF+WM

CLV	Wr	Wf	Wm	R'	F'	M'
۰.۴۷۹۱۶۵۰۴۹	۰	۰.۲۰۱۴۹۵۱۴۶	۰.۲۷۷۶۶۹۹۰۳	۰	۱	۱
۰.۰۶۷۱۶۸۲۷۷	۰.۰۲۹۶۸۵۶۱۴	۰.۰۶۳۳۴۸۹۷۹	۰	۰.۵۱۲۳۸۳۶۴	۰.۳۱۴۳۹۴۵۶۷	۰
۰.۲۱۳۵۸۲۲۹۶	۰.۰۲۹۶۸۵۶۱۴	۰.۰۸۱۵۰۵۹۰۵	۰.۱۰۲۳۹۰۷۷۷	۰.۳۹۸۲۵۱۸۳۵	۰.۴۰۴۵۰۵۵۴۹	۰.۳۶۸۷۵
۰.۲۷۲۱۴۹۶۳۷	۰.۰۷۰۸۳۸۹۵۷	۰	۰.۲۰۱۳۱۰۶۸	۰.۹۵۰۳۵۰۷۰۶	۰	۰.۷۲۵

Wm=M'*M"
Wf=F'*F"
Wr=R'*R"

m'=(M-Mmin)/(Mmax-Mmin)
F'=(F-Fmin)/(Fmax-Fmin)
R'=(R-Rmin)/(Rmax-Rmin)



شکل۷ : خروجی تحلیلی ٤ خوشه اصلی با استفاده از روش K-Means

Figure 7: Analytical output of 4 main clusters using K-Means method

خوشه بندی مکان‌های پر بازدید از مهم ترین گام های اصلی در ایجاد و برقراری سامانه ی مدیریت گردشگری با گردشگران است. در هرم مکان‌های پر بازدید چهار بخش داده‌های شناسایی گردیده است که عبارتد از:

- مکان‌های تفریحی – گردشگری
- مکان‌های فرهنگی و تاریخی
- مکان‌های مذهبی و زیارتی
- مکان‌های اجتماعی
- مکان‌های علمی و اقتصادی سهم بسیار اندکی را در بازدیدها به خود اختصاص داده اند.

هرم مکان‌های پربازدید در بخش الگوی سفر مسافران شهر تهران بر مبنای گردشگری ابزاری به منظور سوق دادن گردشگران به گردشگری می باشد و انتخاب مکان مورد نظر را به خود گردشگر واگذار می نماید، به عبارت دیگر از منظر هرم مکان‌های پربازدید توسط گردشگر مهم این است که گردشگران مکان هارا برحسب علایق و میزان رضایت‌مندی خود با کمترین هزینه و بیشترین رضایت ثبت نماید. با توجه به ایجاد چهار خوشه در مرحله خوشه بندی مکان ها مطابق با هرم زیر اقدام به نام گذاری خوشه‌ها می گردد(شکل۸):

شکل ۸ – هرم بازدید از مکان ها

Figure 8 : Pyramid of visiting places

به این ترتیب نام گذاری در هرم بازدید از مکان‌های تناوبی مطابق با جدول زیر صورت می پذیرد:

جدول ۱۳: جدول تعداد هر خوشه و CLV آنها

Table13: The table of the number of each cluster and their CLV

عنوان	تعداد گردشگر	میانگین CLVe
خوشه اول	۲۷	۰.۴۷۹۱۶۵۰۴۹
خوشه دوم	۱۹	۰.۲۱۳۵۸۲۲۹۶
خوشه سوم	۴۴	۰.۲۷۲۱۴۹۶۳۷
خوشه چهارم	۱۳	۰.۰۶۷۱۶۸۲۷۷

جدول ١٤ : جدول نام گذاری هرم براساس شاخص CLV

Table 14: Pyramid naming table based on CLV index

عدد clv	نام خوشه در هرم
۰.۵> - ۰.۴=<	پلاتینی
۰.۴>- ۰.۳=<	طلایی
۰.۳> - ۰.۱۵=<	نقره‌ای
۰.۱۵>	مسی

جدول ۱۵: نام گذاری خوشه ها در هرم بازدید از مکان ها

Table15: Naming the clusters in the pyramid of visiting places

فراوانی هر دسته	عدد clv خوشه	نام خوشه در هرم	خوشه	نوع مکان
۲۷	۰.۴۷۹۱۶۵۰۴۹	پلاتینی	خوشه اول	مکان‌های تفریحی
۱۹	۰.۲۷۲۱۴۹۶۳۷	نقره‌ای	خوشه دوم	اجتماعی
۴۴	۰.۲۱۳۵۸۲۲۹۶	نقره‌ای	خوشه سوم	مذهبی و زیارتی
۱۳	۰.۰۶۷۱۶۸۲۷۷	مسی	خوشه چهارم	مکان‌های فرهنگی و تاریخی

با مشخص شدن میانگین CLVe هر خوشه، امکان تشکیل هرم مکان‌های پربازدید بر مبنای داده‌های داوطلبانه توسط گردشگر در اپلیکیشن تلفن همراه فراهم گردیده و این هرم با چهار بخش پلاتینی طلایی، نقره‌ای و مسی تشکیل گردید.

نتیجه گیری

بر اساس نتایج این پژوهش در بررسی ۳۰٤ مکان مورد بازدید۱۰۳ مکان بدون تکرار توسط گردشگران به چهار گروه بر اساس متغیرهای R و F و M تقسیم گردیده است و بر اساس هرم ارزش‌گذاری مکان ها گروه‌های پلاتینی، طلایی، نقره‌ای و مسی تشکیل گردید. ارزش گذاری ها براساس سطح بیشترین بازدید در شهر تهران تعیین شده است.

جدول ۱۶: تفسیر خوشه ها بر اساس مدل FRM

Table16: Interpretation of clusters based on the FRM model

عنوان	تعداد	نرمال M	نرمال F	نرمال R	CLVe مرکز خوشه
وزن	۱۰۳	۰.۳۶	۰.۵۵	۰.۰۹	-
خوشه اول	۲۷	ضعیف	ضعیف	قوی	۰.۴۷۹۱۶۵۰۴۹
خوشه دوم	۱۹	ضعیف	ضعیف	قوی	۰.۰۶۷۱۶۸۲۷۷
خوشه سوم	۴۴	ضعیف	ضعیف	قوی	۰.۲۱۳۵۸۲۲۹۶
خوشه چهارم	۱۳	متوسط	متوسط	ضعیف	۰.۲۷۲۱۴۹۶۳۷

تفسیر خوشه اول: مکان‌های قرار گرفته در این خوشه، مدت زمان زیادی از آخرین بازدید آن گذشته، تکرار بازدید آن کم بوده و مدت زمان بازدید از آنها بسیار زیاد بوده است. بنابراین سطح بازدید خوبی دارند.

تفسیر خوشه دوم: مکان‌های قرار گرفته در این خوشه، مدت زمان زیادی از آخرین بازدید آن گذشته، تکرار بازدید آن کم بوده و مدت زمان بازدید از آنها بسیار زیاد بوده است.

تفسیر خوشه سوم: مکان‌های قرار گرفته در این خوشه، مدت زمان زیادی از آخرین بازدید آن گذشته، تکرار بازدید آن کم بوده و مدت زمان بازدید از آنها طولانی بوده است.

تفسیر خوشه چهارم: مکان‌های قرار گرفته در این خوشه، مدت زمان زیادی از آخرین بازدید آن نگذشته است، تکرار بازدید آن متوسط بوده و مدت زمان بازدید از آنها کم بوده است. هریک از خوشه ها دارای ۳ مشخصه اصلی R و F و M می باشند و براساس این ۳ متغیر اصلی مکان های مورد بازدید بررسی شده اند.

در واقع مشخصه تازگی مکان ها R : فاصله زمانی بین آخرین مکان ثبت شده توسط گردشگر تا پایان دوره ثبت اطلاعات در اپلیکیشن است. تکرار مکان‌های مورد بازدید F : تعداد مکان‌هایی است که یک گردشگر در یک دورهٔ زمانی خاص سفرکرده است. مدت زمان بازدید از مکان ها M : مدت زمان بازدیدهای صورت گرفته توسط گردشگر.

جدول ۱۷: تفسیر خوشه ها بر اساس مدل CLV

Table 17: Interpretation of clusters based on the CLV model

فراوانی هر دسته	عدد clv خوشه	نام خوشه در هرم	خوشه	نوع مکان
۲۷	۰.۴۷۹۱۶۵۰۴۹	پلاتینی	خوشه اول	مکان‌های تفریحی
۱۹	۰.۲۷۲۱۴۹۶۳۷	نقره‌ای	خوشه دوم	اجتماعی
۴۴	۰.۲۱۳۵۸۲۲۹۶	نقره‌ای	خوشه سوم	مذهبی و زیارتی
۱۳	۰.۰۶۷۱۶۸۲۷۷	مسی	خوشه چهارم	مکان‌های فرهنگی و تاریخی

خوشه اول: در گروه مکان‌های پلاتینی قرار دارند این گروه شامل۲۷ مکان می باشند و ٤٠ بازدید در بازه زمانی این پژوهش ثبت شده است. این مکان ها براساس شاخصCLV بیشترین بازدید هارا دارا بوده اند.

خوشه دوم : در گروه مکان‌های نقره‌ای قرار دارند این گروه شامل ۱۹ مکان می باشند و ۳۸ بازدید در بازه زمانی این پژوهش ثبت شده است.این گروه از مکان ها بازدید کمتری نسبت به خوشه اول داشته اند.

خوشه سوم: در گروه مکان‌های نقره‌ای قرار دارند این گروه شامل ٤٤ مکان می باشند و ١٤٦ بازدید در بازه زمانی این پژوهش ثبت شده است.

خوشه چهارم: در گروه مکان‌های مسی قرار دارند این گروه شامل ۱۳ مکان می باشند و ۷۰ بازدید در بازه زمانی این پژوهش ثبت شده است.

با توجه به مثبت بودن هر سه متغیر R و F و M در این پژوهش CLVe محاسبه شده عددی مثبت می‌باشد و به عبارت دیگر هرمکان که از CLVe بالاتری برخوردار باشد برای گردشگری ارزشمند تر است. در بخش بندی مکان‌های مورد بازدید که از چهار بخش اصلی مکان‌های مذهبی، تفریحی، فرهنگی و تاریخی و مکان‌های علمی تشکیل می گردد ابتدا الویت انتخاب گردشگر در بازدید از مکان ها، بخش تفریحی سپس بخش فرهنگی-تاریخی پس از آن مذهبی و زیارتی و در بخش بعدی اجتماعی بوده است. با اینکه در وزن دهیAHP ، مکان‌های مذهبی بیشترین امتیاز را دریافت نموده اند اما زمانی که متغیرهای RFM در داده کاوی اثرگذار شدند و وزن آنها در داده‌های نرمال ضرب شد به ترتیب مکان‌های ذکر شده فوق اولویت یافتند.

بنابراین مکان‌های تفریحی و مکان‌های فرهنگی وتاریخی بیشترین الویت را در بازدیدهای صورت گرفته توسط گردشگران استان تهران در بازه زمانی پژوهش را دارا بوده اند، استفاده شده است که نوع مکان ها در خوشه بندی

مشخص گردد. ناگفته نماند که هر مکان دارای یک ID مختص به خود است که در خوشه بندی ID هر مکان استفاده شده است و علت استفاده از ID مکان به این دلیل بوده است که شناسایی نوع مکان‌های موردبازدید در داده کاوی صورت پذیرد.

پیشنهادات پژوهش

- تحقیق صورت گرفته در منطقه گردشگری استان تهران انجام شده و انجام تحقیق در سایر استان ها می‌تواند در ارایه نقشه راه برای گردشگران موثر واقع شود.
- با توجه به نرخ رشد تغییرات فرهنگ، سازه‌های شهری و.... تحقیق در بازه‌های زمانی دیگر نیز انجام و نرخ تغییرات میل به سفر (افزایش و یا کاهش) و علل آن بررسی شود.
- محققان میتوانند از روش‌های داده کاوی و آماری مانند هوش مصنوعی نسبت به استخراج نتایج اقدام و نتایج حاصله را مقایسه و تحلیل نمایند.
- اپلیکیشن، نرم افزار داده کاوی و... در تحقیق به صورت شفاف بیان شده و محققین می‌توانند در تحقیقات آتی از سایر روش ها در طراحی اپلیکیشن، سرور، گستره داوطلبان شرکت کننده در ثبت داده‌های داوطلبانه مکانی هوشمند استفاده نمایند.
- روش‌های ارزیابی اطلاعات داوطلبانه و اینکه چه روشی در پژوهش بر اساس هدف پژوهشگر کاربرد دارد، در تحقیقات آینده بررسی گردد.

Refrences

- Babaei, E (2015). Designing a web service for flood relief management with the help of voluntary geographic information based on open source technology. Journal of Spatial Analysis of Environmental Hazards, Volume 3, Number 4, pp. 1-12. [doi:10.18869/acadpub.jsaeh.3.4.1].
- Sohrabi, Babak and Iraj, Hamida (2014). The book of big data management in private and public sectors. Samet Publications, pp. 25-27.
- Vahedi Targhabeh, Behzad et al. (2014). Quantitative estimation of the practical quality of voluntary spatial information with the help of fuzzy linguistic quantifiers and multi-criteria decision maker OWA Scientific Research Journal of Mapping Sciences and Techniques, Volume 3, Number 4, pp. 63-65. https://www.sid.ir/paper/249536/en.
- Yazdi,E et all (2013). Zoning of geotourism potentials of Iran. Proceedings of the first geotourism conference, Tehran, volume 5, number 2, pp. 28-39. https://www.researchgate.net/publication/352002349.
- Pashaei, Z&Malik, M. (2014). Investigating the quality of voluntary spatial information from the perspective of the index with emphasis on the appropriateness index. Promotional Scientific Journal of Mapping and Spatial Information, 9, 73-76. https://www.researchgate.net/profile/Mohammad-Malek/publication/324149521_brrsy_kyfyt_dadh_hay_mkany_dawtlbanh_az_mnzr_shakhs_ba_takyd_br_shakhs_brazndgy_astfadh/links/5ac116c245851584fa759b10/brrsy-kyfyt-dadh-hay-mkany-dawtlbanh-az-mnzr-shakhs-ba-takyd-br-shakhs-brazndgy-astfadh.pdf.
- Mahalati, Soroush (2018). Tourism and tourist attractions. The first Iranian Studies and Tourism onference, Journal of Tourism and Travel, Volume 8, Number 5, pp. 27. Doi:10.22034/JTD.2020.241178.2093.
- Vahidnia, Mohammad Hossein and colleagues (2016). Mass distribution of geographic information resources by voluntary users of devices with the aim of rapid relief. Scientific-research journal of mapping sciences and techniques, volume 7, number 3, pages 161-175.
- Ameri, Mohammad and colleagues (2018). Providing a model to attract people's participation in sustainable land transportation development projects. Journal of Environmental Science and Technology, Volume 13, Number 2, pp. 67-79.
- Mohammadi, Nazila and Malik, Mohammad Reza (2012). Geospatial information environments of people: characteristics and challenges. Promotional Scientific Journal of Mapping and Spatial Information, Volume 24, Number 2, pp. 55-48.
- Nazimi, Farzad (2015). geographical information system. Proceedings of the Iranian Geography Association, volume 5, number 6, pp. 58-71.
- Ushahidi Okolloh (2018). Web2 tools for crowdsourcing crisis information. Participatory learning and action, E-Commerce Article, Volume 7 , Number 5 :pp 65-70. DOI: 10.4236/ce.2013.47A2010.
- Azizkhani, M & Malek, M.R. (2016). Evaluation of precision of volunteer geographic information in Haitian earthquakes. International Conference on Integrated Natural Disaster Management Volume 4 Number 4 : pp 128-140 .
- Ching-Hsue, Chen (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with applications , Data Mining Application Article, Volume 5, Number 1: pp 4176-4184.
- Suliu , Daniel (2004). Voluntary data collection and evaluation of the results of different environmental impact assessment methods. Article Environmental Assessment, Volume 12 Number 1 : pp 198-210 .
- Goodchild, M. (2010). "Crowdsourcing geographic information for disaster response: a research frontier.". International Journal of Digital Earth, Volume 4 Number 4 : pp 231-241. DOI:10.1080/17538941003759255.
- Hansen, ch. (2011). InGeospatial semantics and the semantic web. International Journal of Spatial Data Infrastructures Research, 2 , pp 73-96 .

- Hollenstein, Purves (2010). Exploring place through user-generated content: using flickr to describe city cores. Journal of Spatial Information Science, Volume 1 Number 3 : pp 21- 48 . DOI:10.5311/JOSIS.2010.1.3.
- Keiningham, Timothy (2006). Approches to measurement and management of customer value, . Journal of relationship marketing, Volume 5 Number 2 : pp 37-54 . https://doi.org/10.1300/J366v05n02_03.
- Klonner at all (2016). Geographic Information in Natural Hazard Analysis: A Systematic Literature Review of Current Approaches with a Focus on Preparedness and Mitigation. ISPRS International Jornal of Geo- Information , Volume 12 Number 5 : pp 88-108 .
- Madadipouya , ali (2015). A New Decision tree method for Data mining in Medicine. Advanced Computational Intelligence: An International Journal (ACII), Volume 2 Number 8 : pp 23-48 . https://airccse.org/journal/acii/papers/2315acii04.pdf.
- Ngai, Chau (2009). Application of data mining techniques incustomer relationship management: A literature review and classification. Expert Systems with Applications, Journal E-Commerce, Volume 36 Number 2 : pp 90-101 .
- Ngai, Chau (2009). Application of data mining techniques incustomer relationship management: A literature review and classification. Expert Systems with Applications, Journal E-Commerce, Volume 36 Number 2 : pp 90-101 .
- Scheider Sumeil et al. (2016). Semantic referencing ofgeosensor data and volunteered geographic information. InGeospatial semantics and the semantic web, Journal volunteered geographic information, Volume 2 Number 3 : pp 88-96 . DOI:10.1007/978-1-4419-9446-2_2.
- Senaratne , Heli (2016). A review of volunteered geographic information quality assessment methods. Interntional Journal of Geographical Information Science, Volume 11 Number 6 : pp 139 -157 .
- Shukla, Parmis (2015). Application of Data Mining in Customer Relationships Management (CRM). International Journal of Technology Management & Humanities, Volume 1, Issue 1, : pp 4123-4140 .
- Farzanyar , z & Cercone, N. (2016). Trip pattern mining using large scale geo-tagged photos. Proceedings of the International Conference on Computer and Information Science and Technology Ottawa, Ontario, Canada, pp 210-230. https://avestia.com/CIST2015_Proceedings/papers/113.pdf.
- Kafashpour colleagues (2013),Article on segmentation of customers based on their lifetime value using in Johashhai Journal of Public Management - 5th year, 15th issue, Spring 2013pp: 63-84 .

Data mining of passenger travel patterns using voluntary spatial information (case study: Tehran)

** Ensiyeh Mihanparast*

Master's degree, Remote Sensing and Geographical Information System, Khwarazmi University, Tehran, Iran (corresponding author). ensiyeh.m1995@gmail.com

Javad Sadidi

Assistant Professor, Department of Remote Sensing and Geographical Information System, Faculty of Geographical Sciences, Kharazmi University, Tehran, Iran, jsadidi@gmail.com

Extended abstract

Introduction:

Investigating the most visited places is one of the most important issues in the way of voluntary geographic data in order to investigate the travel pattern of tourists. Travel pattern recognition (pattern recognition) is a branch of machine learning. It can be said that pattern recognition is receiving raw data and making decisions based on data classification. Pattern recognition can be defined as the classification of input data into known classes by extracting important features. Based on this, people provide the system with their information that has been collected for a specific purpose. In this attitude, each person can be like a sensor that observes the environment around him. One of the main motivations for moving towards such processes is access to a large amount of information and low cost of accessing accurate sources of geographic information. Therefore, the aim of the research is to provide a conceptual model that can be used to correctly store the data entered into the voluntary database and evaluate the travel pattern of tourists in Tehran based on the voluntary information.

and Malek, (2014) Pashaei, in his research titled "Investigation of the quality of voluntary spatial information from the perspective of the index, with emphasis on the appropriateness index", have come to the conclusion that voluntary spatial information is a type of information that is voluntarily provided by ordinary people, without the need for Scientific education is prepared according to local knowledge. Voluntary spatial data has many advantages compared to standard spatial data, including easy information access and fast updating. But the quality of these data is one of the important discussions in this direction. Pashaei and Malek, (2013) in the study of the spatial information environment of the people of Gostar (features and challenges) have come to the conclusion that the importance of voluntary spatial data is based on the fact that People, especially the residents of a place, have the most knowledge about describing their lives, and generally these people are without mapping knowledge. Unlike standard spatial data that use metadata.

Materials and Methods:

Data analysis (data mining) is a multi-stage process during which the data provided through the use of collection tools in the statistical sample (community); They are summarized, coded and categorized and finally processed in order to provide the basis for various analyzes and connections between these data in order to test the hypotheses. In this process, the data are refined both conceptually and experimentally, and various statistical techniques play a significant role in the conclusions and generalizations. Khaki, 2008)) in this section, the collected data has been evaluated. The development of technology and the emergence of new facilities in the field of the Internet has provided a platform for the production of spatial data by the general public and voluntarily. In this way, unlike the

traditional process of data generation, spatial data is produced by each user and made available to others for free. This phenomenon with the name of spatial information has had a significant impact on the production and sharing of spatial data, so that every person can be both a producer and a user of spatial data. VGI is a rich and valuable source of spatial data that allows users to visualize the world based on their perception and perspective. The data of the present research was extracted from the SQL database which was stored using the voluntary information of the users. For the purpose of this research, about 300 samples were extracted from this database and provided to the researcher.

Findings and Discussion:

Clustering of the most visited places is one of the most important main steps in creating and establishing a tourism management system with tourists. In the pyramid of the most visited places, four data sections have been identified, which are:

- Recreational - tourism places
- Cultural and historical places
- Religious and pilgrimage places
- Social places
- Scientific and economic places have a very small share in visits.

The pyramid of the most visited places in the travel pattern of travelers of Tehran is a tool to drive tourists to tourism and leaves the choice of the desired place to the tourist. Tourists can register the places according to their interests and level of satisfaction with the lowest cost and the highest satisfaction. With the determination of the average CLVe of each cluster, it is possible to form a pyramid of the most visited places based on voluntary data provided by tourists in the mobile phone application, and this pyramid has four platinum sections. Gold, silver and copper were formed.

Conclusion:

Based on the results of this research, in the survey of 304 places visited by tourists, 103 places without repetition were divided into four groups based on R, F, and M variables, and based on the place valuation pyramid, platinum, gold, silver, and copper groups were formed. Valuations are determined based on the most visited level in Tehran.

Interpretation of the first cluster: the places located in this cluster, a long time has passed since the last visit, the repetition of the visit is low, and the duration of their visit is very long. Therefore, they have a good level of visits.

Interpretation of the second cluster: the places located in this cluster, a long time has passed since the last visit, the repetition of the visit is low, and the duration of their visit is very long.

Interpretation of the third cluster: the places located in this cluster, a long time has passed since the last visit, the repetition of the visit is low, and the duration of their visit is long.

Interpretation of the fourth cluster: the places located in this cluster, not much time has passed since its last visit, the repetition of its visit was average, and the duration of its visit was short. Each of the clusters has 3 main characteristics R, F and M and based on these 3 main variables the visited places have been checked.

In fact, the characteristics of the newness of the places: R is the time interval between the last place registered by the tourist and the end of the information registration period in the application. Repetition of visited places: F is the number of places that a tourist has traveled in a certain period of time. Duration of visiting places: M duration of visits made by tourists.

The first cluster: in the group of platinum places, this group includes 27 places and 40 visits were recorded in the time frame of this research. These places have the most visits based on the CLV index.

The second cluster: There are silver places in the group. This group includes 19 places and 38 visits have been recorded in the time frame of this research. This group of places has had fewer visits than the first cluster.

The third cluster: in the group of silver places, this group includes 44 places and 146 visits were recorded in the time frame of this research.

The fourth cluster: in the group of copper places, this group includes 13 places and 70 visits were recorded in the time frame of this research.

Considering the positiveness of all three variables R, F and M in this research, the calculated CLVe is a positive number, and in other words, any place with a higher CLVe is more valuable for tourism. In the division of the visited places, which consists of four main parts: religious, recreational, cultural and historical places and scientific places, the priority of the tourist's choice in visiting the places is the recreational part, then the cultural-historical part, then the religious and pilgrimage part, and in The next part was social. Although in AHP weighting, religious places have received the most points, but when the RFM variables were effective in data mining and their weight was multiplied by normal data, the above mentioned places were prioritized.

Therefore, recreational places and cultural and historical places have had the highest priority in the visits made by tourists of Tehran province in the research period, it has been used to determine the type of places in clustering. It goes without saying that each place has its own ID, which is used in the ID clustering of each place, and the reason for using the place ID was to identify the type of places visited in data mining.

Keywords: Voluntary geographic information - travel pattern - data mining - tourism - RFM method