Users clustering Based on Search Behavior Analysis Using the LRFM Model (Case Study: Iran Scientific Information Database (Ganj))

Somayeh Fatahi*
PhD in Computer Engineering; Assistant Professor; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: fatahi@irandoc.ac.ir

Mohammad Rabiei
PhD in IT Engineering; Assistant Professor; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: m.rabiei@irandoc.ac.ir

Iranian Journal of Information Processing and Management

Received: 21, Apr. 2020 **Accepted: 30, Jun. 2020**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 36 | No. 2 | pp. 419-442
Winter 2021

Abstract: Iran scientific information database (Ganj) which includes almost one million scientific records provides the search opportunity in dissertations, domestic scientific journals, articles, conferences, research projects, and governmental reports. A large number of researchers meet the needs of their scientific and research resources from Ganj database daily. Users' needs and behaviors are variant and understanding it helps system administrators to use different strategies to manage the better databases and provide efficient services to users. One way to understand users' needs is to cluster them based on their behavior and identify the features of each cluster. This study aims to cluster the users based on the analysis of their search behavior using the LRFM model. In this study, the search log data of Ganj users were collected for three months, the LRFM attributes were calculated, and then the K-means algorithm was applied to them. The optimal number of clusters was calculated based on different criteria. Based on customer value matrix the results of customer clustering users in four groups are efficient, suspicious, unreliable, and intermittent and based on customer loyalty Marcus users are categorized in loyal, potential, insecure and newcomers.

Keywords: Clustering, LRFM Model, Marcus Customer Value Matrix, User Behavior Analysis, Ganj Database

* Corresponding Author

استفاده از مدل LRFM برای خوشه‌بندی کاربران بر اساس تحلیل رفتار جست‌وجو

(مورد مطالعه: پایگاه اطلاعات علمی ایران «گنج»)

سمیه فتاحی
دکتری مهندسی کامپیوتر، هوش مصنوعی و رباتیک؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
پدیدآور رابط Fatahi@irandoc.ac.ir

محمد ربیعی
دکتری مهندسی فناوری اطلاعات، تجارت الکترونیک؛ استادیار؛ پژوهشکدهٔ فناوری اطلاعات، گروه پژوهشی کسب‌وکار الکترونیک؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
M.Rabiei@irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۹/۰۲/۰۲ **پذیرش:** ۱۳۹۹/۰٤/۱۰ مقاله برای اصلاح به مدت ۲۸ روز نزد پدیدآوران بوده است.

چکیده: پایگاه اطلاعاتی «گنج» پژوهشگاه علوم و فناوری اطلاعات با برخورداری از نزدیک به یک میلیون رکورد علمی امکان جست‌وجو در پایان‌نامه‌ها، نشریات علمی داخلی، مقالات، همایش‌ها، طرح‌های پژوهشی و گزارش‌های دولتی را فراهم می‌کند. روزانه تعداد زیادی از پژوهشگران نیازهای منابع علمی و پژوهشی خود را از پایگاه «گنج» تأمین می‌کنند. نیازها و رفتارهای کاربران مختلف این پایگاه متنوع بوده و شناخت دقیق‌تر آن موجب خواهد شد که مدیران این پایگاه بتوانند استراتژی‌های متناسب با هر یک از گروه‌های کاربران را به‌منظور مدیریت بهتر پایگاه و ارائهٔ خدمات کاراتر اتخاذ نمایند. یکی از راه‌های شناخت کاربران، خوشه‌بندی آن‌ها و شناخت ویژگی‌های هر خوشه است. هدف این پژوهش، خوشه‌بندی کاربران بر اساس تحلیل رفتار جست‌وجوی آن‌ها با استفاده از مدل LRFM است. در این پژوهش، داده‌های لاگ جست‌وجوی کاربران پایگاه «گنج» به مدت سه ماه جمع‌آوری و مورد استفاده قرار گرفت. با استفاده از داده‌های لاگ رفتار جست‌وجوی کاربران، شاخص‌های مدل LRFM محاسبه و سپس، الگوریتم K-means بر روی آن‌ها اعمال و تعداد خوشهٔ بهینه بر اساس معیارهای مختلف محاسبه شد. نتایج به‌دست‌آمده از خوشه‌بندی بر اساس ماتریس ارزش مشتری، کاربران را در چهار گروه بهره‌مند مشکوک،

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳٦ | شمارهٔ ۲ | ص‌ص ٤١٩-٤٤٢
زمستان ۱۳۹۹

نامطمئن و متناوب قرار می‌دهد و بر اساس ماتریس وفاداری، کاربران در چهار گروه وفادار، بالقوه، نامطمئن و تازه‌وارد ارزیابی می‌شوند.

کلیدواژه‌ها: خوشه‌ندی، مدل LRFM، ماتریس ارزش مشتری مارکوس، تحلیل رفتار کاربران، پایگاه اطلاعاتی گنج

۱. مقدمه

با توجه به رشد فراوان صفحات وب و استفادهٔ روزافزون کاربران از موتورهای جست‌وجو و پایگاه‌های علمی به‌منظور بازیابی اطلاعات، توجه به رفتار جست‌وجوی کاربران اهمیت بالایی دارد. بازیابی اطلاعات و تحلیل رفتار کاربران دو جزء مورد توجه در موتور جست‌وجوی یک پایگاه اطلاعات علمی هستند که مطالعهٔ آن‌ها می‌تواند به بهبود عملکرد پایگاه کمک کند. تحقیقات در زمینهٔ موتورهای جست‌وجو، بهبود الگوریتم‌های آن‌ها و افزایش رضایت کاربران همواره مورد توجه محققان بوده است (Hernon and Altman 1996; Hawking et al. 2001; Agichtein, Brill and Dumais 2006; Huo, Zhao and Ren 2017). در سال‌های اخیر، ارزیابی موتورهای جست‌وجوی بومی بسیار مورد توجه قرار گرفته و محققان سعی دارند با تحلیل آن‌ها از منظرهای مختلف، عملکرد آن‌ها را بهینه نمایند (شعله و همکاران ۱۳۹۵؛ عباس‌پور ۱۳۸۵؛ عظیم‌زاده، فرهادی و اثنی‌عشری ۱۳۹۴).

بسیاری از موتورهای جست‌وجو داده‌های زیادی دربارهٔ کاربران جمع‌آوری و ذخیره می‌کنند، ولی توانایی کشف دانش پنهان و باارزش در این داده‌ها را ندارند. در نتیجه، این داده‌ها هیچ‌وقت به دانش کاربردی تبدیل نمی‌شوند. سازمان‌ها میل دارند اطلاعات ناشناخته، معتبر و قابل درک از بانک‌های اطلاعاتی عظیم خود را استخراج کرده و از این اطلاعات برای افزایش کاربران و کسب سود بیشتر استفاده کنند. یکی از روش‌های مؤثر برای ارتقای رتبهٔ هر پایگاه اطلاعاتی حفظ کاربران کنونی و جذب کاربران جدید است. افزایش رتبه، موجب روانه شدن ورودی‌ها می‌شود. در نتیجه، به اعتماد بیشتر کاربران و باور آن‌ها مبنی بر این‌که این پایگاه، پایگاه علمی قوی و بزرگی است، کمک می‌کند. کاربران وفادار سرمایه‌های پایگاه داده‌های علمی هستند. بنابراین، سامانه برای حفظ این کاربران بایستی تمامی تلاش خود را به‌کار برد (ایمانی و عباسی ۱۳۹۶). نکتهٔ قابل توجه این است که اکثر شرکت‌ها به‌راحتی ۲۵ درصد مشتریان خود را از دست می‌دهند و هزینهٔ جذب یک مشتری جدید پنج برابر هزینهٔ حفظ یک مشتری قدیمی است (Kolter 1994).

مدیران پایگاه‌های داده‌های علمی زمانی می‌توانند به موفقیت امید داشته باشند که بتوانند رفتارهای آیندهٔ کاربران را پیش‌بینی کنند. برای این منظور سازمان‌ها باید داده‌هایی کامل از فعالیت‌های گذشتهٔ کاربران خود داشته باشند. هدف اصلی از ایجاد پروفایل برای هر کاربر آگاهی از فعالیت‌های گذشتهٔ کاربران است تا بتوان با استفاده از روش‌های داده‌کاوی رفتار جست‌وجوی کاربران را تحلیل نمود. در صورت تمرکز بر ایجاد ارزش برای کاربر می‌توان وفاداری وی را به استفاده از پایگاه دادهٔ علمی افزایش داد. بنابراین، حفظ و وفادار ساختن کاربران یکی از مهم‌ترین و حیاتی‌ترین راهبردهای هر پایگاه دادهٔ علمی و موتور جست‌وجوست.

اتخاذ سیاست مناسب برای حفظ و مدیریت کاربران نیازمند استفاده از مدلی است که قادر باشد مناسب‌ترین خدمت را به انواع کاربران پیشنهاد دهد (یوسفی‌زاد و ثریایی ۱۳۹۷). از آنجا که تعداد و تنوع کاربران در استفاده از پایگاه اطلاعات علمی بسیار زیاد است و هر یک از کاربران الگوهای رفتاری خاص خود را دارد، از رویکرد داده‌کاوی می‌توان به‌عنوان ابزاری برای ایجاد یک مدل هوشمند در سیاست‌گذاری برای حفظ کاربران وفادار و جذب کاربران جدید استفاده کرد تا مدیران بتوانند با خوشه‌بندی کاربران بر اساس شاخص‌های مناسب و برچسب‌گذاری آن‌ها با استفاده از دیگر منابع اطلاعاتی و اهداف بالادستی خود، سیاستی مناسب برای هر خوشه را پیشنهاد دهند. از این رو، سؤال اساسی پژوهش حاضر برای پاسخگویی این است که کاربران «گنج» را از نظر رفتاری به چند دسته می‌توان تقسیم کرد و ویژگی‌های هر یک از این دسته‌ها به‌منظور اتخاذ استراتژی مناسب برای مدیریت کاربران این پایگاه چیست؟

پایگاه اطلاعات علمی ایران «گنج»، یکی از مهم‌ترین پایگاه‌های علمی و تنها مرجع رسمی گردآوری و اشاعهٔ پایان‌نامه‌ها و رساله‌های تحصیلات تکمیلی کشور است. این پایگاه دارای بیش از یک میلیون رکورد علمی است که بیش از نیمی از آن شامل پایان‌نامه‌ها و رساله‌های تحصیلات تکمیلی است (Rabiei, Hosseini-Motlagh and Minaei Bidgoli 2019). به‌روزرسانی پایگاه، بهبود موتورجست‌وجو و رابط کاربری، و تغییر سیاست‌های دسترسی کاربران به محتوای پایگاه اطلاعاتی موجب شده است که اخیراً تغییرات چشمگیری در میزان جست‌وجو و رفتار جست‌وجوی کاربران «گنج» مشاهده شود (Rabiei, Hosseini-Motlagh and Haeri 2017; Fatahi and Ershadi 2020). از این رو، شناسایی گروه‌های مختلف کاربران این پایگاه به‌منظور آشنایی با اهداف و نیازهای هر

گروه و اعمال سیاست‌گذاری صحیح برای هر یک از گروه‌ها ضروری است. برای این منظور داده‌های لاگ رفتار جست‌وجوی کاربران به‌مدت سه ماه جمع‌آوری و پس از انجام پیش‌پردازش بر روی آن‌ها، شاخص‌های مدل LRFM، از آن استخراج شد. سپس، کاربران، با اجرای الگوریتم K-means خوشه‌بندی و بر اساس ماتریس ارزش «مارکوس»، دسته‌بندی شده و برای هر دسته تحلیل انجام شده است.

در ادامهٔ این مقاله مروری بر ادبیات این حوزه خواهیم داشت. سپس، مبانی نظری بررسی خواهد شد. روش پژوهش و نتایج در بخش‌های بعدی ارائه می‌شوند.

۲. مرور ادبیات

در تحلیل لاگ کاربران در زمان استفاده از موتورهای جست‌وجو از سال ۱۹۹۵ به بعد مطالعات گسترده‌ای انجام شده است. «ولفرام» لاگ کاربران را از چهار سیستم بازیابی اطلاعات متفاوت جمع‌آوری و رفتار کاربران و مشخصه‌های جست‌وجوی آنان را مورد مطالعه قرار داد (Wolfram 2008). «پارک، لی و بای» داده‌های لاگ یک موتور جست‌وجوی کره‌ای را مورد مطالعه قرار دادند. در این پژوهش مشخص شد که کاربران مشغول رفتارهای ساده مانند پرس‌وجوهای کوتاه هستند و به‌ندرت از ویژگی‌های پیشرفتهٔ پایگاه استفاده می‌کنند و در نهایت، تعداد کمی از صفحات نتیجه را مشاهده می‌کنند (Park, Lee and Bae 2005). یافته‌های «پارک و لی» بر روی لاگ کاربران کتابخانهٔ دیجیتال ملی کشور کُره نشان می‌دهد که کاربران تمایل دارند کمترین زمان و تلاش را برای جست‌وجوی اطلاعات در آن پایگاه صرف نمایند. از سوی دیگر، توقف کاربران در لاگ کتابخانهٔ دیجیتال طولانی‌تر از موتورهای جست‌وجوی وب است (Park and Lee 2013). «استنمارک» با استفاده از روش خوشه‌بندی K میانگین خوشه‌های کاربران یک شبکهٔ اینترانت را تعیین نمود (Stenmark 2008). «کاتوریا» و همکارانش با استفاده از روش خوشه‌بندی K میانگین، ۱۳۰ هزار پرس‌وجوی موتورهای جست‌وجو را خوشه‌بندی نمودند. پرس‌وجوها به سه دستهٔ اطلاعاتی، هدایت‌گر و مبادلاتی تقسیم‌بندی شده‌اند. یافته‌ها نشان می‌دهد که ۷۵ درصد پرس‌وجوی‌های وب، اطلاعاتی است و این در حالی است که ۱۲ درصد جست‌وجوها برای هدایت به سایتی خاص و انجام مبادله مانند بانکداری اینترنتی یا خرید کالا هستند (Kathuria et al. 2010). پژوهشی دیگر، فاکتورهای انسانی مرتبط با رفتار کاربران از جمله سبک‌های شناختی، سطح مهارت و تفاوت‌های جنسیتی را بررسی کرده و سه

الگوریتم خوشه‌بندی را برای درک رفتار کاربران به‌کار برده است (Frias-Martinez et al. 2007).

یک هفته از اطلاعات لاگ کاربران یک پایگاه جست‌وجوی عکس شامل پرس‌وجوها، دنبالهٔ کلیک‌های کاربر، صفحات مشاهده‌شده و رفتار کاربر در هر صفحه مورد تحلیل قرار گرفته است. نتایج این پژوهش نشان داده است که رفتار آن دسته از کاربرانی که به‌دنبال یافتن عکس در پایگاه‌های عکس هستند، با کاربران جست‌وجوی معمولی در وب متفاوت است (Zhijing Wu et al. 2017).

افزون بر پژوهش‌های ذکرشده، پژوهش‌هایی نیز بر روی تحلیل لاگ رفتار کاربران پایگاه اطلاعاتی «گنج» صورت پذیرفته است. از آن جمله می‌توان به پژوهشی که «خسروی و جمالی مهموئی» انجام داده و رفتار جست‌وجوی کاربران از منظر زمان اتصال، تعداد جست‌وجو، تعداد کلمات مورد جست‌وجو، میزان استفاده از جست‌وجوی پیشرفته و نظایر آن را بررسی کرده‌اند، اشاره کرد (۱۳۹۳). برخی از پژوهش‌ها تا حدی وارد فضای تحلیل محتوای جست‌وجوی کاربران شده و علاوه بر توصیف شکل و ساختار جست‌وجوها، کلیدواژه‌های مورد جست‌وجوی کاربران را نیز رده‌بندی نموده و با استفاده از نظر خبرگان برچسب‌زنی کرده‌اند (فتاحی و نعیمی صدیق ۱۳۹۵). در پژوهشی که «ربیعی» و همکاران با به‌کارگیری لاگ جست‌وجوی کاربران «گنج» انجام داده‌اند، تعداد ۱۰۶,۳۱۴ رکورد مربوط به جست‌وجوهای انجام‌شده در حوزهٔ محیط زیست مورد تحلیل قرار گرفته و میزان جست‌وجوی هر یک از زیرشاخه‌های محیط زیست و تعداد پاسخ موتور جست‌وجو به آن‌ها مبنای تحلیل بوده و در نهایت، زیرشاخه‌ها به چهار گروه «جست‌وجوی زیاد و پاسخ زیاد»، «جست‌وجوی زیاد و پاسخ کم»، «جست‌وجوی کم و پاسخ زیاد» و «جست‌وجوی کم و پاسخ کم» دسته‌بندی شده‌اند (Rabiei, Hosseini-Motlagh and Haeri 2017). همچنین، اطلاعات لاگ سه ماه از جست‌وجوهای پایگاه «گنج» به‌منظور استخراج ارتباطات معنایی که کاربر در فرایند اصلاح جست‌وجوهای خود مورد استفاده قرار می‌دهد، تحلیل شده است. بررسی رفتار کاربران در فرایند جست‌وجو و با استفاده از ابزار اصطلاحنامه و روش تحلیل معنایی نشان داده است که کاربران برای یافتن نتیجهٔ مطلوب در جست‌وجو، متن مورد کاوش را بر اساس روابط معنایی اصلاح می‌کنند و می‌توان از این روابط در بهبود طراحی موتور جست‌وجوی پایگاه استفاده کرد (Karimi, Babaee and Hosseini Beheshti 2018).

قابل ذکر است که پژوهش‌هایی نیز در به‌کارگیری مدل RFM و LRFM در دسته‌بندی و خوشه‌بندی مشتریان انجام شده است. به‌عنوان مثال، «علیزاده ذوارم و کریمی مزیدی» روشی سیستماتیک برای تحلیل ویژگی‌های رفتاری خرید مشتریان ارائه دادند که از طریق آن بتوانند کارایی مدیریت رابطه با مشتری را بهبود بخشند. آن‌ها از مدل LRFM و ابزار خوشه‌بندی استفاده کردند (Alizadeh Zoeram, Karimi Mazidi 2018). در سال ۱۳۹۷، «یوسفی‌زاد و ثریایی» به‌دنبال ارائهٔ الگویی برای ارائهٔ خدمات کلیدی به مشتریان بودند و در این راستا از مدل RFM استفاده کردند و بر اساس آن کاربران را با استفاده از دو الگوریتم DBSCAN و K-means خوشه‌بندی نمودند (یوسفی‌زاد و ثریایی ۱۳۹۷). در پژوهشی که توسط «بازدارو بهرامی» انجام گرفت، ارزش عمر فعلی مشتریان بر اساس مدل RFM توسعه یافته و با استفاده از وزن‌دهی سلسله‌مراتبی تعیین شد. سپس، رفتار مشتریان تحلیل و در نهایت، ارزش طول عمر هر مشتری که شامل ارزش عمر فعلی و آیندهٔ مشتری است، تعیین شد (بازدار و بهرامی ۱۳۹۷).

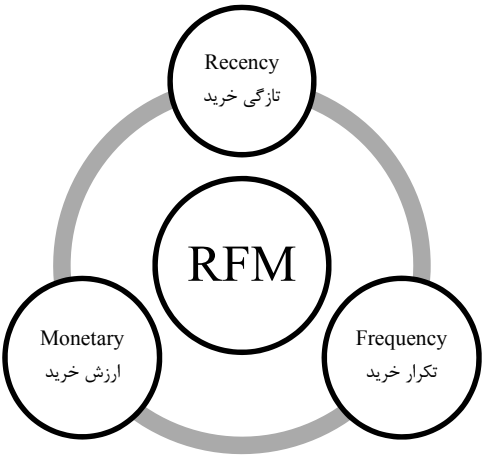
با بررسی پژوهش‌های پیشین معلوم شد که آن دسته از پژوهش‌هایی که به تحلیل رفتار جست‌وجوی کاربران در پایگاه «گنج» پرداخته‌اند، صرفاً به توصیف نوع کاربری و رفتار کلی کاربران معطوف بوده و هیچ‌کدام از آن‌ها با هدف خوشه‌بندی کاربران از منظر رفتارهای مختلف جست‌وجوگری به تحلیل اطلاعات نپرداخته‌اند. پژوهش حاضر پس از بررسی و تحلیل رفتار کاربران با استفاده از مدل توسعه‌یافته RFM یعنی LRFM و به‌کارگیری روش‌های داده‌کاوی به شناسایی و تحلیل گروه‌های کاربری پرداخته است.

۳. مبانی نظری

۳-۱. ابزار بازاریابی

مدل RFM[1]، یک مدل رایج و معتبر در مدیریت ارتباط با مشتری است که آن را اولین‌بار Hughes (1994) معرفی کرد. در این مدل از سه شاخص مربوط به داده‌های مبادلاتی مشتریان برای تحلیل رفتار و ارزش‌گذاری آن‌ها استفاده می‌شود. شاخص‌ها به شرح زیر است (شکل ۱):

1. recency, frequency monetary



شکل ۱. مدل RFM

تازگی[1]: این شاخص به فاصلهٔ زمانی بین آخرین فعالیت خرید کاربر تا کنون اشاره می‌کند. کمتر بودن این فاصله نشانگر بالاتر بودن ارزش این شاخص در مدل است.

تکرار[2]: نشان‌دهندهٔ تعداد تراکنش‌های کاربر در یک دورهٔ زمانی مشخص است. بیشتر بودن تکرار، نشانگر بالاتر بودن ارزش این شاخص در مدل است.

ارزش[3]: نشان‌دهندهٔ ارزشی است که برای مبادلات در یک دورهٔ زمانی مشخص صرف شده است. بیشتر بودن این شاخص، بیانگر ارزش بالاتر این شاخص در مدل است.

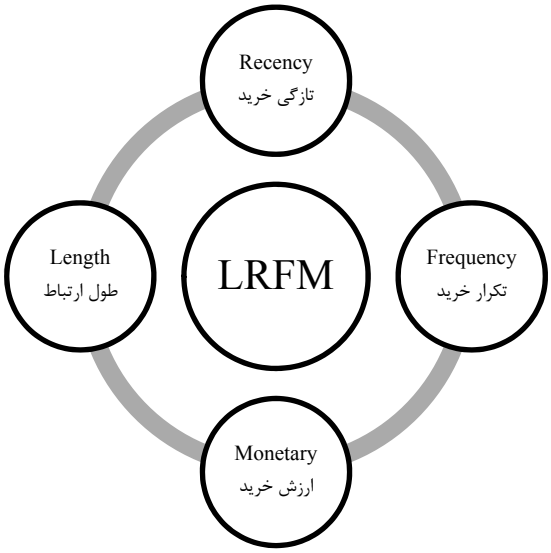
نظرات مختلفی پیرامون وزن هر یک از شاخص‌ها در مدل وجود دارد. اما از دید «هوگس» این سه شاخص اهمیت یکسانی دارند (Hughes 1994). برای خوشه‌بندی کاربران با استفاده از مدل RFM، مقادیر شاخص مدل برای هر کاربر محاسبه می‌شود. بنابراین، کاربران بر اساس مقادیر سه شاخص خوشه‌بندی می‌شوند. فرض اساسی مدل این است که الگوهای آیندهٔ مبادله و خرید مشتری همانند الگوهای گذشته و حال است. از ویژگی‌های مدل، راحتی محاسبه و قابل درک بودن و توانایی RFM در پیش‌بینی رفتار آینده و ارزیابی وفاداری مشتری است. با استفاده از مدل RFM، می‌توان سرعت پاسخگویی به مشتریان را افزایش داد و تعداد مشتریان بیشتری را حفظ نمود که این کار خود باعث

1. recency 2. frequency 3. monetary

افزایش درآمد می‌شود. این مدل همچنین می‌تواند با تحلیل اطلاعات گذشتهٔ مشتریان برای برنامه‌ریزی اهداف آینده بسیار مؤثر باشد.

بر اساس این مدل، هرچه F و R بیشتر باشد، احتمال آن‌که تراکنش جدیدی با مشتری صورت بگیرد، بیشتر است و همچنین، اگر M نیز بزرگ‌تر باشد، احتمال بازگشت مشتری برای خرید بیشتر است. در مدل RFM فرض بر این است که مشتریانی که در هر یک از متغیرهای مدل ارزش بالاتری دارند، بهترین مشتریان هستند.

تلاش‌های متعددی در زمینهٔ بهبود شاخص‌های مدل RFM، صورت گرفته است. یکی از مهم‌ترین آن‌ها مدل LRFM است (Chang and Tsay 2004). این مدل علاوه بر سه متغیر تازگی، تکرار و ارزش، طول مدت همکاری کاربران را نیز لحاظ می‌کند (شکل ۲).



شکل ۲. مدل LRFM

هدف از توسعهٔ مدل RFM به LRFM این بود که مدل RFM نمی‌تواند مشتریان دارای ارتباط بلندمدت و مشتریان دارای ارتباط کوتاه‌مدت با سازمان را مشخص نماید. طول ارتباط مشتری می‌تواند بر وفاداری و سودآوری مشتری تأثیرگذار باشد و افزایش طول ارتباط با مشتری، وفاداری مشتری را بهبود خواهد بخشید. در واقع، در مدل RFM مشتریانی که به‌تازگی ارزش مالی بالایی برای سازمان ایجاد کرده و در کوتاه‌مدت دارای تناوب خریدی بیش از متوسط تناوب خرید در بین مشتریانی باشند که تکرار خرید داشته‌اند،

به‌عنوان مشتریان با ارزش انتخاب می‌شوند، در حالی که عامل طول ارتباط با سازمان نادیده گرفته شده است.

عمدهٔ کاربران پایگاه اطلاعاتی «گنج» دانشجویان و پژوهشگران تحصیلات تکمیلی هستند (Rabiei, Hosseini-Motlagh and Minaei Bidgoli 2019) به‌دلیل این‌که «گنج» یک پایگاه علمی بوده و شامل منابع پژوهشی و به‌ویژه پایان‌نامه‌ها و رساله‌های تحصیلات تکمیلی است، این کاربران بیشتر در زمان تحصیل خود از این پایگاه استفاه می‌کنند و پس از آن به‌ندرت به آن مراجعه می‌کنند. از این رو، مفهوم وفاداری در مورد این کاربران صرفاً در مدت‌زمان تحصیل و در یک بازهٔ ۲ تا ۴ ساله صادق است. همچنین، از آنجا که خدمات پایگاه اطلاعاتی «گنج» به‌صورت رایگان ارائه می‌شود، شاخص ارزش به میزان فعالیت کاربر و دریافت خدمات مختلف از این پایگاه معطوف است. به‌عبارت دیگر، کاربری که میزان فعالیت بیشتری در این پایگاه داشته و از منابع موجود در این پایگاه به‌نحوی درست بهره‌برداری می‌کند، مبلغی بابت استفاده از این منابع پرداخت نمی‌کند، اما موجب اعتبار بیشتر پایگاه خواهد شد و از همین رو، کاربرِ باارزشی به حساب می‌آید.

۳-۲. ابزار خوشه‌بندی

خوشه‌بندی فرایندی است که به کمک آن می‌توان مجموعه‌ای از اشیا را به گروه‌های مجزا تقسیم‌بندی کرد به‌نحوی که اعضای متعلق به هر گروه که یک خوشه نام دارد، دارای بیشترین ویژگی‌های مشابه باشند و در عوض میزان شباهت بین خوشه‌ها کمترین مقدار باشد. خوشه‌بندی در رشته‌های مختلف از جمله پزشکی (زرنگارنیا و همکاران ۱۳۸۹؛ Khanmohammadi, Adibeig and Shanehbandy 2017)، زمین‌شناسی (شادمان، تخم‌چی و خیراللهی ۱۳۹۱)، علم‌سنجی (ابویی اردکان، عابدی جعفری و آقازاده ۱۳۸۹)، مدیریت مصرف انرژی (Wang et al. 2016) و مسایل اقتصادی (Cai, Le-Khac and Kechadi 2016) کاربرد دارد.

روش‌های مختلفی برای خوشه‌بندی وجود دارد که از این میان روش K-means یکی از رایج‌ترین آن‌هاست. مزایایی چون سادگی پیاده‌سازی، قابلیت مقیاس‌پذیری با مجموعه داده‌های بزرگ، تضمین همگرایی و قابلیت تطابق با داده‌های جدید و کارایی بالا در خوشه‌بندی داده‌های با ابعاد بالا موجب شده است که از این روش در این پژوهش استفاده شود.

در این روش فرض می‌کنیم که مشاهدات (x1,x2,...,xn) دارای d بُعد هستند و قصد داریم آن‌ها را به k خوشه به نام S={S1,S2,...,Sk} تقسیم کنیم. اعضای خوشه‌ها باید به شکلی از مشاهدات انتخاب شوند که تابع مجموع مربعات درون خوشه‌ها[1] کمینه شود. بنابراین، تابع هدف در این الگوریتم به‌صورت زیر نوشته می‌شود.

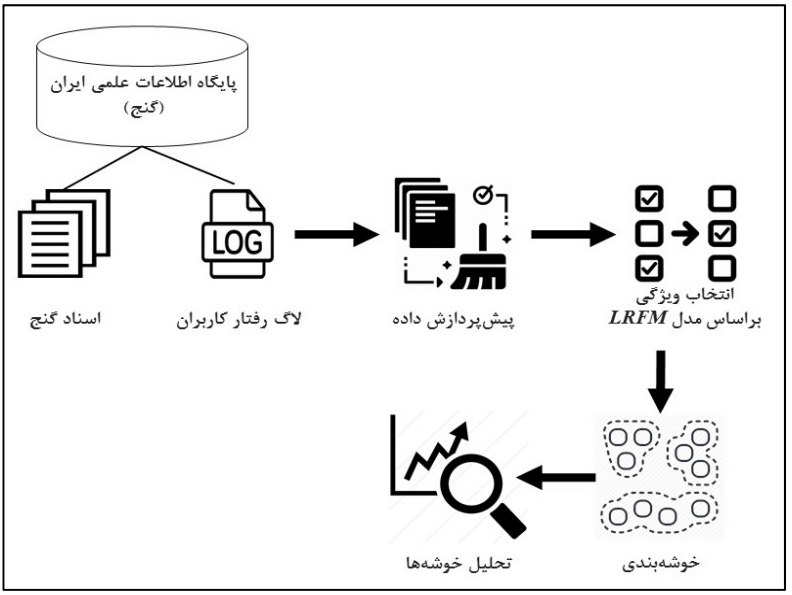
$$ {}_s argmin \sum_{i=1}^{k} \sum_{x \in S_i} ||X - \mu_i||^2 = {}_s argmin \sum_{i=1}^{k} |S_i| Var S_i \qquad (۱) $$

در اینجا منظور از μi میانگین خوشهٔ Si و |Si| تعداد اعضای خوشهٔ i ام است. منظور از argmin کمینه‌کردن فاصله بین اعضاست. در واقع، هدف از خوشه‌بندی این است که فاصله بین اعضای داخل یک خوشه کمینه باشد. فرمول بالا بیانگر محاسبهٔ فاصله و کمینه‌کردن آن است. البته، می‌توان نشان داد که کمینه کردن این مقدار به‌معنای بیشینه‌سازی میانگین مربعات فاصله بین نقاط در خوشه‌های مختلف (Between-Cluster Sum Of Squares- BCSS) است، زیرا طبق قانون واریانس کل، با کم شدن مقدار WCSS، مقدار BCSS افزایش می‌یابد، زیرا واریانس کل ثابت است. WCSS معیار تنوع و تغییرپذیری نمونه‌های داخل یک خوشه است. به‌طور کلی، یک خوشه که دارای مجموع مربعات کمتری است، متراکم‌تر از خوشه‌ای است که مجموع مربعات بزرگ‌تری دارد.

٤. روش پژوهش

این پژوهش قصد دارد با خوشه‌بندی کاربران بر مبنای شاخص‌های مدل LRFM امکان تعریف سیاست‌های مناسب برای مدیریت کاربران را برای پایگاه اطلاعاتی «گنج» فراهم سازد. بنابراین، این مطالعه از نظر نوع پژوهش، کاربردی است. همچنین، با توجه به این‌که این پژوهش با توصیف و تفسیر شرایط و روابط موجود بین کاربران و فعالیت آن‌ها به خوشه‌یابی کاربران پایگاه اطلاعاتی «گنج» می‌پردازد و مجموعهٔ منظمی از داده‌ها را جمع‌آوری می‌کند، مطالعهٔ حاضر از نوع پژوهش توصیفی است. چارچوب پژوهش به‌طور خلاصه در شکل ۳، آمده است.

1. Within-Cluster Sum of Squares (WCSS)

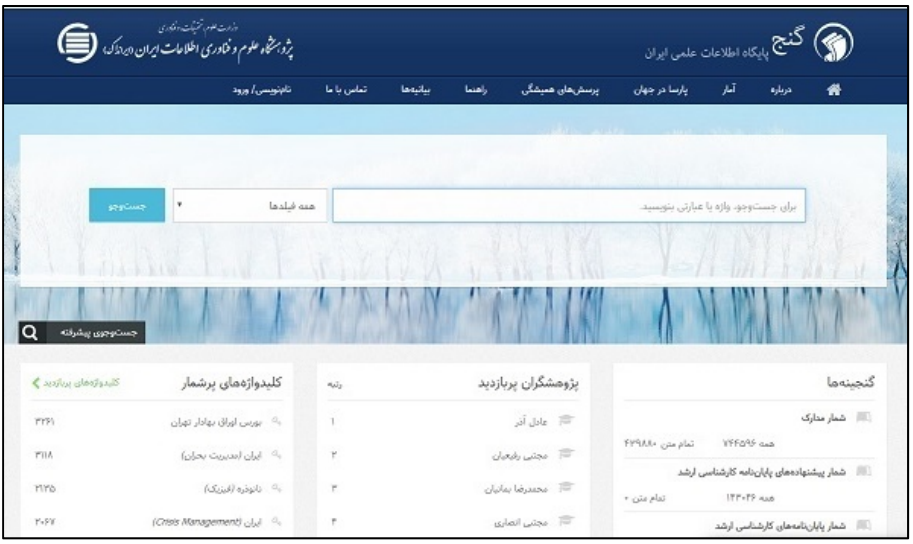


شکل ۳. چارچوب پژوهش

همان‌طور که در شکل ۳، آمده، پس از جمع‌آوری لاگ عملکرد کاربران در پایگاه «گنج»، داده‌ها مورد پیش‌پردازش قرار گرفته و ویژگی‌های رفتاری کاربران انتخاب خواهد شد. در این مرحله لازم است مقادیر این ویژگی‌ها که از جنس و دامنه‌های مختلف هستند، نرمال‌سازی شوند. سپس، بر اساس معیارهای مختلف تعداد خوشه‌های مناسب انتخاب و با استفاده از ابزارهای خوشه‌بندی در داده‌کاوی، کاربران در تعداد خوشهٔ به‌دست‌آمده در مرحلهٔ قبل افراز خواهند شد. در مرحلهٔ بعدی رفتار مختلف این خوشه‌ها تحلیل خواهد شد و راهکارهای مختلف برای مدیریت هر یک از خوشه‌ها با توجه به رفتار کاربری آن‌ها پیشنهاد خواهد شد.

٤-١. جمع‌آوری داده‌های لاگ جست‌وجو در «گنج»

پایگاه اطلاعاتی علمی ایران «گنج» که توسط «پژوهشگاه علوم و فناوری اطلاعات» راه‌اندازی شده و مدیریت می‌شود، با برخورداری از نزدیک به یک میلیون رکورد علمی، امکان جست‌وجو برای پژوهشگران را در پایان‌نامه‌ها، نشریات علمی داخلی، مقالات، همایش‌ها، طرح‌های پژوهشی و گزارش‌های دولتی فراهم می‌کند (شکل ۴). پژوهش حاضر بر روی لاگ جست‌وجوی کاربران در این پایگاه انجام می‌شود.



شکل ٤. نمایی از پایگاه «گنج»

در این پژوهش ابتدا داده‌های جست‌وجوی سه ماههٔ پایگاه «گنج» از تاریخ ۱۳۹۸/۰۳/۰۵ تا ۱۳۹۸/۰۶/۳۱ جمع‌آوری شده است. داده‌ها شامل ۹،۱۳۰،۰۰۰ رکورد متعلق به ۲۶،۲۸۵ کاربر است. هر لاگ کاربر شامل فیلدهای زیر است: آدرس IP مورد درخواست، تاریخ و زمان فعالیت، تاریخ و زمان به‌روزرسانی، شناسهٔ کاربر[1]، شناسهٔ جلسه[2]، نوع لاگ[3]، URL مورد ارجاع[4]، و لاگ اطلاعات[5]. نوع لاگ شامل ۲۳ مقدار متفاوت است که در جدول ۱، نمایش داده شده است.

جدول ۱. لیست ویژگی‌های موجود در مجموعهٔ داده

نوع لاگ	عنوان	توضیح
۱	Login	ورود کاربر به پایگاه به‌صورت شناخته‌شده
۲	Search	جست‌وجوی معمولی
۳	Show profile	مشاهدهٔ اطلاعات میز کاربری (پروفایل)
۴	Advanced	جست‌وجوی پیشرفته
۵	Click	کلیک بر روی یکی از نتایج جست‌وجو

1. user id
2. session id
3. application log type
4. the referring URL
5. information log

نوع لاگ	عنوان	توضیح
۶	Show document	مشاهدهٔ یک مدرک
۷	Request keywords	درخواست کاربر برای نمایش کلیدواژه‌ها
۸	Request abstract	درخواست کاربر برای نمایش چکیده
۹	Request-list	درخواست کاربر برای نمایش فیلدهای فهرست نوشته‌ها
۱۰	Show-admin	نمایش اطلاعات مربوط به ادمین
۱۱	Show_Doc_No	نمایشگر مدرک
۱۲	Fulltext	نمایش تمام‌متن مدرک
۱۳	Show-research	نمایش گراف پدیدآوران
۱۴	Research	آراِس‌اِس برای جست‌وجوی معمولی
۱۵	Export	دریافت خروجی برای مدارک انتخابی
۱۶	Share	اشتراک‌گذاری مدرک در شبکه‌های اجتماعی
۱۷	Tagfeed	آراِس‌اِس برای یک کلیدواژه
۱۸	Subjects-feed	آراِس‌اِس برای یک موضوع
۱۹	Researchers-feed	آراِس‌اِس برای یک پژوهشگر
۲۰	Fields-feed	آراِس‌اِس برای یک رشتهٔ تحصیلی
۲۱	Errorr-ports	ثبت گزارش اشکال برای یک مدرک
۲۲	Advance-feed	آراِس‌اِس برای جست‌وجوی پیشرفته
۲۳	Organizations-feed	آراِس‌اِس برای یک سازمان

٤-۲. پیش‌پردازش داده‌ها

از آنجا که دادهٔ مربوط به فیلد اطلاعات به‌صورت رشته ذخیره شده، در بعضی رکوردها این رشته به‌درستی ذخیره نشده است. بنابراین، در مرحلهٔ پیش‌پردازش، فیلد اطلاعات همهٔ رکوردها بررسی و در صورت نیاز مقدار فیلد به‌روزرسانی می‌شود. برای این‌که ردگیری و تحلیل فعالیت‌های مستمر کاربران امکان‌پذیر باشد، رکوردهای مربوط به کاربر مهمان[1] حذف شده است. در نهایت، ۱،۰۴۸،۵۷۵ رکورد برای فرایند داده‌کاوی در نظر گرفته شده است.

1. guest user

۵. خوشه‌بندی کاربران

۵-۱. انتخاب ویژگی بر اساس مدل LRFM

با توجه به این‌که هدف از این پژوهش به‌کارگیری مدل LRFM است، شاخص‌های این مدل با ویژگی‌های لاگ جست‌وجو انطباق داده شده و مقداردهی شده‌اند (جدول ۲).

جدول ۲. تعریف شاخص‌های مدل LRFM

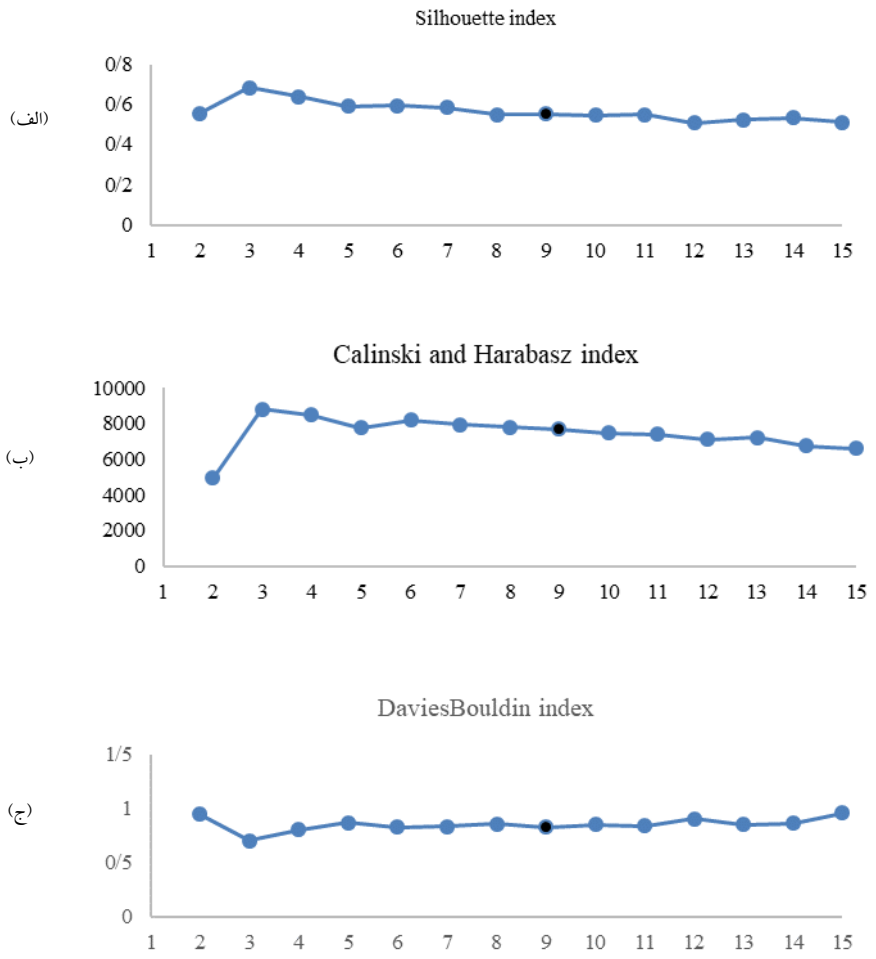
شاخص	تعریف
تازگی (R)	فاصلهٔ زمانی آخرین لاگین تا کنون به دقیقه
تکرار (F)	مجموع تعداد جست‌وجوی ساده و پیشرفته
ارزش (M)	مجموع تعداد کلیک بر روی یکی از نتایج جست‌وجو، مشاهدهٔ یک مدرک، درخواست نمایش کلیدواژه‌ها، درخواست نمایش چکیده، درخواست نمایش فیلدهای فهرست نوشته‌ها، نمایش چند صفحهٔ اول مدرک و نمایش تمام متن مدرک
طول مدت (L)	مدت‌زمان کل فعالیت کاربر در پایگاه تا کنون به دقیقه

تمامی داده‌های مجموعهٔ داده بر اساس مدل LRFM برای هر یک از شاخص‌ها به‌طور جداگانه و بر اساس فرمول زیر نرمال‌سازی شده‌اند.

$$R_i = \frac{R_i - Min\,(R)}{Max\,(R) - Min\,(R)} \quad , \quad L_i = \frac{L_i - Min\,(L)}{Max\,(L) - Min\,(L)}$$
$$F_i = \frac{F_i - Min\,(F)}{Max\,(F) - Min\,(F)} \quad , \quad M_i = \frac{M_i - Min\,(M)}{Max\,(M) - Min\,(M)} \qquad (۲)$$

۵-۲. خوشه‌بندی با استفاده از روش K-means

در استفاده از روش K-means مهم‌ترین فاکتور انتخاب K مناسب است. تعداد خوشه‌ها را بر اساس کیفیت ساختاری می‌توان با آزمون‌های متفاوتی تعیین نمود. در این مرحله از پژوهش سه معیار Sihouette، CalinskiHarbasz و DaviesBouldin مورد بررسی قرار گرفت (Davies and Bouldin 1979؛ Caliński and Harabasz 1974؛ Rousseeuw 1987). برای مقادیر مختلف K عمل خوشه‌بندی با روش K-means انجام گرفت و مقادیر این معیارها برای K به‌دست آمد (شکل ۵).



شکل ۵. مقایسهٔ معیارهای Silhouette (الف)، CalinskiHarbasz (ب) و DaviesBouldin (ج) برای انتخاب تعداد خوشه‌ها

معیار Sihouette (نیم‌رخ) به پیوستگی[1] درون خوشه‌ها و میزان تفکیک‌پذیری آن‌ها بستگی دارد. مقدار نیم‌رخ برای هر نقطه، میزان تعلق آن را به خوشه‌اش در مقایسه با خوشهٔ مجاور اندازه می‌گیرد. تمرکز این معیار بر کیفیت خوشه‌بندی انجام شده تکیه دارد. در واقع، این معیار مشخص می‌کند که پراکندگی داده‌ها در خوشه‌ها به چه صورت است. هرچه مقدار این معیار بالاتر باشد، کیفیت خوشه‌بندی نیز بالاتر است.

1. cohesion

معیار Davies Bouldin از شباهت بین دو خوشه استفاده می‌کند که بر اساس پراکندگی یک خوشه و عدم شباهت بین دو خوشه تعریف می‌شود. در حقیقت این شاخص، میانگین حداکثر نسبت پراکندگی درون به پراکندگی بین خوشه‌ها را محاسبه می‌کند. معیار CalinskiHarbasz، دو ماتریس پراکندگی درون خوشه‌ای و ماتریس پراکندگی میان‌خوشه‌ای را تعریف می‌کند. بر اساس معیار CalinskiHarbasz، هرچه مقدار این معیار بیشتر باشد، خوشه‌ها متراکم‌تر و از هم تفکیک‌یافته‌تر هستند.

همان‌طور که در شکل ۵، مشاهده می‌شود، با توجه به سه معیار مشخص‌شده انتخاب K=9 انتخاب بهینه برای تعداد خوشه‌های الگوریتم است و انتخاب تعداد خوشهٔ بالاتر تغییر محسوسی در بهبود معیارهای مد نظر ندارد. نتایج خوشه‌بندی و تحلیل آن‌ها در بخش نتایج ارائه می‌شود.

٦. یافته‌ها

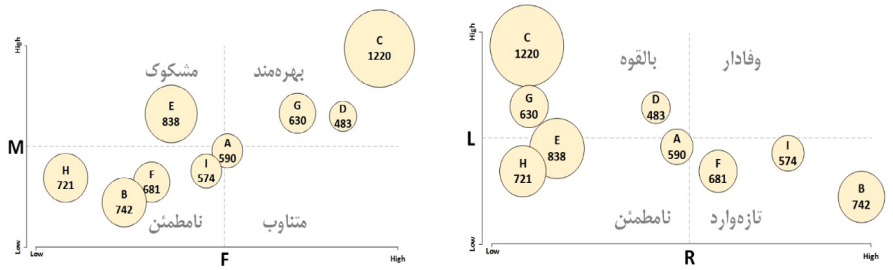
بعد از اجرای روش خوشه‌بندی K-means، ۹ خوشه به‌دست آمد که نتایج آن در جدول ۳، آمده است. مقادیر میانگین هر یک از شاخص‌های LRFM در هر خوشه محاسبه شده است.

جدول ۳. نتایج خوشه‌بندی لاگ کاربران پایگاه «گنج» بر اساس مدل LRFM

خوشه	L	R	F	M	تعداد کاربران
خوشهٔ A	۰/۰۲۴	۰/۳۷	۰/۵۳۹	۰/۰۲۳	۵۹۰
خوشهٔ B	۰/۰۱۱	۰/۷۷۶	۰/۲۵۵	۰/۰۱۰	۷۴۲
خوشهٔ C	۰/۰۴۳	۰/۰۷۱	۰/۹۱	۰/۰۴۲	۱۲۲۰
خوشهٔ D	۰/۰۳۳	۰/۳۴۳	۰/۸۱۱	۰/۰۲۹	۴۸۳
خوشهٔ E	۰/۰۲۲	۰/۱۱۵	۰/۳۸۶	۰/۰۲۴	۸۳۸
خوشهٔ F	۰/۰۱۴	۰/۴۶۸	۰/۲۸۲	۰/۰۱۲	۶۸۱
خوشهٔ G	۰/۰۳۱	۰/۰۹۳	۰/۶۷۵	۰/۰۳۰	۶۳۰
خوشهٔ H	۰/۰۱۳	۰/۰۷۳	۰/۱۳۴	۰/۰۱۴	۷۲۱
خوشهٔ I	۰/۰۲۱	۰/۰۶۲	۰/۵۱۴	۰/۰۲۱	۵۷۴

«مارکوس» در سال ۱۹۹۸، یک ماتریس با عنوان ماتریس ارزش با ابعاد فرکانس خرید (F) و ارزش (M) و یک ماتریس به‌عنوان ماتریس وفاداری مشتری با ابعاد طول

ارتباط مشتری (L) و زمان معاملهٔ اخیر (R) پیشنهاد کرد (Marcus 1998). دو ماتریس وفاداری کاربران شکل ۶ (الف) و ارزش کاربران شکل ۶ (ب) با استفاده از نه خوشه در ادامه آمده است. در این ماتریس تعداد رکوردهای هر خوشه با اندازهٔ هر دایره نشان داده شده و بسته به بالا و پایین بودن مقادیر، هر یک از شاخص‌های LRFM از دو منظر وفاداری و ارزش کاربران این خوشه‌ها در موقعیت‌های چهارگانه هر ماتریس قرار گرفته‌اند.



(الف) (ب)

شکل ٦. ماتریس وفاداری کاربران (الف) و ماتریس ارزش کاربران (ب)

در ماتریس وفاداری کاربران شکل ۶ (الف)، کاربران وفادار کسانی هستند که L و R بالایی دارند و کاربران بالقوه کسانی هستند که L بالا و R پایین دارند. L و R پایین متعلق به کاربران نامطمئن و L پایین و R بالا متعلق به کاربران تازه‌وارد است. بر اساس ماتریس ارزش مشتری «مارکوس»، بهترین مشتریان در قسمتی قرار می‌گیرند که مقدار F و M بالا باشد. در دسته‌بندی کاربران «گنج»، ما این دسته از کاربران را کاربران بهره‌مند نامیدیم. در عوض کاربران با حجم دانلود بالا که دارای M بالا و F پایین هستند، کاربران مشکوک هستند؛ چرا که میزان دانلود آن‌ها با میزان جست‌وجوی آن‌ها متناسب نیست. به‌عبارت دیگر، به نظر نمی‌رسد که به‌دنبال منبعی در یک موضوع خاص باشند و تنها با یک جست‌وجوی عمومی تعداد قابل توجهی از رکوردهای بازیابی‌شده را دانلود می‌کنند. کاربران نامطمئن کسانی هستند که F و M پایینی دارند و کاربران متناوب به کسانی گفته می‌شود که M پایین و F بالا دارند.

با توجه به این‌که خدمات مختلف پایگاه «گنج» به‌عنوان M در نظر گرفته شده است، طبیعی است که کاربرانی که دارای فرکانس بالای تعامل با پایگاه باشند، از خدمات آن نیز استفاده می‌کنند. از این رو، در ماتریس ارزش کاربران (شکل ۶-ب)، پایگاه «گنج»

تقریباً فاقد کاربران متناوب است و این بدان‌معناست که کاربرانی که به‌صورت مداوم به پایگاه مراجعه دارند، از خدمات آن نیز بهره‌مند می‌شوند و یا کاربرانی که از خدمات مناسبی بهره‌مند نمی‌شوند، به‌طور مکرر به پایگاه مراجعه نمی‌کنند. از سوی دیگر، از آنجا که عمدهٔ این کاربران دانشجویان تحصیلات تکمیلی هستند که در مراحل انجام پایان‌نامه یا رسالهٔ خود از منابع موجود در «گنج» استفاده می‌کنند، عمدتاً پس از این مرحله، مراجعهٔ چندانی به پایگاه ندارند. از این رو، تقریباً فاقد گروه کاربران وفاداری است که به‌صورت مکرر از پایگاه بازدید می‌کنند.

خوشهٔ A کاربران معمولی پایگاه را شامل می‌شود. این کاربران از نظر شاخص‌های میزان جست‌وجو، دانلود، مراجعه به پایگاه و حضور در پایگاه «گنج» در وضعیت میانه قرار دارند و رفتار کاملاً رایجی را از خود نشان می‌دهند. خوشه‌های B، F و I، هم از نظر معیارهای وفاداری و هم از حیث معیارهای تعیین‌کنندهٔ ارزش کاربر در یک گروه قرار می‌گیرند. کاربران این سه خوشه را می‌توان به‌عنوان کاربران تازه‌وارد نامطمئن قلمداد کرد. عمدهٔ فعالیت‌های این کاربران اخیراً انجام شده است. از طرف دیگر، میزان جست‌وجو و دانلود یا مشاهدهٔ متن کامل رکوردها توسط این کاربران بسیار پایین است. اقدام برای افزایش سواد اطلاعاتی این خوشه‌ها و ارائهٔ راهنماهای لازم برای آموزش کاربران می‌تواند موجب جذب بیشتر این خوشه‌ها گردد.

رفتار خوشه‌های C، D و G نیز بسیار به یکدیگر شبیه است و این خوشه‌ها نیز از نظر معیارهای وفاداری و ارزش کاربر در همان گروه هستند. کاربران این خوشه‌ها زمان قابل ملاحظه‌ای را در این پایگاه صرف می‌کنند و از محتوای این پایگاه نیز به میزان زیادی بهره‌مند هستند. این کاربران به خوبی از امکانات جست‌وجوی پایگاه استفاده می‌کنند و در یافتن منابع مورد نیاز خود تبحر دارند.

خوشهٔ E کاربرانی را شامل می‌شود که رفتار آن‌ها با رفتار پژوهشگرانی که انتظار می‌رود مخاطبان اصلی پایگاه «گنج» باشند، متفاوت است. این گروه از کاربران مدت زمان قابل ملاحظه‌ای را در پایگاه صرف می‌کنند و میزان دانلود و استفادهٔ آن‌ها از منابع پایگاه نیز قابل ملاحظه است. با این حال، میزان جست‌وجوی این کاربران کم است. به‌عبارت دیگر، به نظر می‌رسد که پس از یک جست‌وجوی ساده، تعداد زیادی از فایل‌های بازیابی‌شده را دانلود می‌کنند و این احتمال وجود دارد که هدف این گروه از حجم بالای دانلود مطالب پایگاه استفادهٔ شخصی در موضوعات پژوهشی نباشد. در بررسی

دقیق‌تر این خوشه، در برخی موارد کاربرانی به چشم می‌خورند که تقریباً تمام مدت شبانه‌روز در این پایگاه اطلاعاتی مشغول جست‌وجو هستند. از این رو، مدیران پایگاه می‌توانند با بررسی دقیق‌تر کاربران این خوشه و اعمال سیاست‌های مناسب مانع از سو استفادهٔ احتمالی این گروه شوند.

در نهایت، خوشهٔ H، کاربران نامطمئن را تشکیل می‌دهد. این گروه از کاربران زمان اندکی را در پایگاه به‌سر برده، کمترین میزان جست‌وجو را داشته، و به کمترین میزان از منابع موجود در «گنج» استفاده کرده‌اند. این دسته از کاربران احتمالاً کمترین میزان رضایت را از پایگاه داشته باشند که علت آن می‌تواند مواردی چون دشواری تعامل با پایگاه، نبودِ منابع مورد نیاز آن‌ها در محتوای پایگاه و یا ضعف سواد اطلاعاتی آن‌ها برای جست‌وجو و استفاده از قابلیت‌های پایگاه باشد. مدیران پایگاه با انجام اقداماتی مانند نظرسنجی از این کاربران می‌توانند مشکلات یا محدودیت‌های احتمالی پایگاه را درک و در بهبود و ارتقای آن بکوشند.

۷. نتیجه‌گیری

تحلیل رفتار یک راه مفید برای شناخت ویژگی‌های کاربران است که می‌تواند مورد استفادهٔ مدیریت پایگاه اطلاعاتی قرار گیرد و آنان را در سیاست‌گذاری بهتر و انجام اقدامات لازم برای بهبود عملکرد پایگاه و تقویت آن یاری رساند. در این پژوهش از لاگ کاربران پایگاه اطلاعاتی «گنج» برای تحلیل رفتار و خوشه‌بندی آنان استفاده شد. رفتار کاربران بر اساس کدهای مختلف در لاگ پایگاه جمع‌آوری شده است و فعالیت‌هایی مانند جست‌وجوی ساده و پیشرفته، مشاهده و دانلود فایل، مشاهدهٔ چکیده و چند صفحهٔ اول فایل و غیره از جمله فعالیت‌های کاربر است که در تحلیل رفتار او در پایگاه اطلاعاتی مورد استفاده قرار گرفته است. همچنین، سابقهٔ فعالیت کاربر در پایگاه از نظر استمرار و طول مدت حضور کاربر در پایگاه از دیگر مواردی است که در تحلیل رفتار او کاربرد دارد.

پس از جمع‌آوری و پیش‌پردازش داده‌ها، ویژگی‌ها بر اساس معیارهای مدل LRFM در فرایند خوشه‌بندی مورد استفاده قرار گرفت. تعداد خوشه‌های بهینه بر اساس معیارهای مختلف ارزیابی شد و در نهایت، کاربران در ۹ خوشهٔ مختلف قرار گرفتند. این خوشه‌ها بر اساس ماتریس‌های وفاداری و ارزش کاربران تجزیه و تحلیل شده و ویژگی‌های هر

یک از خوشه‌ها بر اساس موقعیت آن‌ها در هر یک از ماتریس‌ها مورد بررسی قرار گرفت. گروهی از کاربران که دارای رفتار مشکوک بوده و حجم قابل توجهی از منابع را تنها با تعداد کمی از جست‌وجوها دانلود می‌کنند، بهتر است تحت نظارت بیشتر قرار گرفته و رفتار آن‌ها رصد شود. همچنین، پیشنهاد می‌شود فرم‌های نظرسنجی در مورد محتوا، ساختار و ابزارهای مورد استفاده در پایگاه در اختیار کاربران نامطمئن قرار گیرد تا مشکلات احتمالی پایگاه از نظر این دسته از کاربران مشخص شده و امکان بهبود آن برای مدیران فراهم آید.

نتایج این پژوهش می‌تواند مورد استفادهٔ مدیران پایگاه اطلاعات علمی قرار گیرد تا با شناسایی و تحلیل رفتار گروه‌های مختلف کاربران، درک عمیق‌تری از نیازها و گرایش‌های آن‌ها داشته باشند و بر مبنای هر یک از گروه‌های مختلف بتوانند سیاست‌های مختلفی را اعمال نمایند. همچنین، با استفاده از نتایج این پژوهش امکان جلوگیری از سوء استفاده‌های احتمالی کاربران و نیز بهبود عملکرد پایگاه اطلاعاتی و شناخت مشکلات احتمالی آن فراهم خواهد شد. تحلیل حجم گسترده‌تری از لاگ اطلاعاتی پایگاه و ارائهٔ استراتژی‌هایی برای بهبود عملکرد آن، ارائهٔ مدل‌های پیش‌بینی رفتار کاربران از طریق تحلیل لاگ‌های پیشین فعالیت آن‌ها، دسته‌بندی و اعتباردهی به کاربران بر مبنای میزان و نوع فعالیت‌آن‌ها از جمله مواردی است که می‌تواند در پژوهش‌های آینده مورد توجه قرار گیرد.

فهرست منابع

ابویی اردکان، محمد، حسن عابدی جعفری، و فتاح آقازاده. ۱۳۸۹. کاربرد روش‌های خوشه‌بندی در ترسیم نقشه‌های علم: مورد کاوی نقشهٔ علم مدیریت شهری. *فصلنامه علوم و فناوری اطلاعات* ۲۵ (۳): ۳۴۷-۳۷۱.

ایمانی، عبدالمجید، و میثم عباسی. ۱۳۹۶. خوشه‌بندی مشتریان بر مبنای مدل RFM با استفاده از الگوریتم C-means فازی (مورد مطالعه: فروشگاه زنجیره‌ای رفاه شهر زاهدان). *پژوهش‌های مدیریت عمومی* ۱۰ (۳۷): ۲۵۱-۲۷۶.

بازدار، علی‌اصغر، و شیرین بهرامی. ۱۳۹۷. پیش‌بینی ارزش عمر مشتری توسط مدل RFM توسعه‌یافته (مطالعه موردی شرکت بیمه). *مهندسی و مدیریت کیفیت* ۸ (۴): ۳۲۷-۳۳۶.

خسروی مریم، و حمیدرضا جمالی مهموئی. ۱۳۹۳. تحلیل لاگ پایگاه اطلاعات و مدارک علمی ایران (ایرانداک) و رفتار جست‌وجوی کاربران آن. *پژوهشنامه پردازش و مدیریت اطلاعات* ۲۹ (۴): ۹۷۹-۱۰۰۶.

زرنگارنیا یلدا، حمید علوی‌مجد، مصطفی رضایی طاویرانی، نصیبه خیر، و علی‌اکبر خادم معبودی. ۱۳۸۹.

کاربـرد خوشـه‌بندی فـازی در تحلیـل پروتئین‌هـای مرتبـط بـا سـرطان‌های مـری، معـده و کلـون بـر اسـاس تشـابهات تفسـیر هستی‌شناسـی ژنـی. *مجلـه علمـی دانشـگاه علـوم پزشـکی سـمنان* ۱۲ (۳۷): ۱۴-۲۱.

شـادمان، محسـن، بهـزاد تخم‌چـی، و حسـن خیراللهـی. ۱۳۹۱. کاربـرد خوشـه‌بندی در تهیـهٔ نقشـه‌های شـبه زمین‌شناسـی بـا اسـتفاده از داده‌هـای ژئوفیزیـک هوابـردی. *نشـریه علمـی پژوهشـی مهندسـی معـدن* ۷ (۱۶): ۱-۱۲.

شـعله، فرزانـه، معصومـه عظیـم‌زاده، محمدمهـدی یدالهـی، اکبـر میرزایـی، و مـژگان فرهـودی. ۱۳۹۵. ارزیابـی خـودکار جویشـگرهای متنـی مبتنـی بـر تجمیـع آرا در حـوزهٔ وب فارسـی. دومیـن کنفرانـس بین‌المللـی وب‌پژوهـی. دانشـگاه علـم و فرهنـگ، تهـران.

عباس‌پـور، جـواد. ۱۳۸۵. ارزیابـی رابـط کاربـر پایـگاه اطلاعـات چکیـده پایان‌نامه‌هـای مرکـز اطلاعـات و مـدارک علمـی ایـران. پایان‌نامـهٔ کارشناسـی ارشـد. دانشـگاه تربیـت مـدرس. تهـران.

عظیـم‌زاده، معصومـه، نویـد فرهـادی، و محمدمهـدی اثنی‌عشـری. ۱۳۹۴. تحلیـل رضایـت کاربـر مبتنـی بـر رفتـار ضمنـی در حیـن جسـت‌وجو. اولیـن کنفرانـس بین‌المللـی وب‌پژوهـی. دانشـگاه علـم و فرهنـگ، تهـران.

فتاحـی، سـمیه، و علـی نعیمـی صدیـق. ۱۳۹۵. تحلیـل رفتـار اطلاع‌یابـی پژوهشـگران در موتـور جسـت‌وجوی سـامانهٔ ملّـی اطلاعـات پایان‌نامه‌هـا/ رسـاله‌های دانش‌آموختـگان داخـل کشـور (گنـج). *فصلنامـه علمـی پژوهشـی مدیریـت اطلاعـات* ۵ (۲): ۳۱-۵۸.

یوسـفی‌زاد، امیـر، و علـی ثریایـی. ۱۳۹۷. بررسـی و خوشـه‌بندی مشـتریان، بـر اسـاس مـدل RFM و طراحـی الگویـی بـرای ارائـهٔ خدمـات بـه مشـتریان کلیـدی. *پژوهشـنامه مدیریـت اجرایـی* ۱۰ (۲۰): ۱۷۵-۱۹۸.

References

Agichtein, E., E. Brill, and S. Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 19-26. Seattle Washington USA.

Alizadeh Zoeram, A., and A. R. Karimi Mazidi. 2018. New Approach for Customer Clustering by Integrating the LRFM Model and Fuzzy Inference System. *Iranian Journal of Management Studies* 11 (2): 351-378.

Cai, F., N. A. Le-Khac, and T. Kechadi. 2016. Clustering approaches for financial data analysis: a survey. arXiv preprint arXiv:1609.08520.

Caliński, T., and J. Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3 (1): 1-27.

Chang, H. H., and S. F. Tsay. 2004. Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation. *Journal of Information Management* 11: 161-203

Davies, D. L., and D. W. Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* 1 (2): 224-227.

Fatahi, S., & N. Ghasem-Aghaee. 2010. An effective intelligent educational model using agent with personality and emotional filters. In Proceedings of the World Congress on Engineering (Vol. 1, pp. 142-147).

Fatahi, S., and M. J. Ershadi. 2020. Assessment of User Satisfaction of Research Theses and Theses in Iranian Scientific Database (Treasure): Based on E-Qual Model. *Iranian Journal of Information*

processing and Management 35 (2): 399-424.

Frias-Martinez, E., S. Y. Chen, R. D. Macredie, and X. Liu. 2007. The role of human factors in stereotyping behavior and perception of digital library users: a robust clustering approach. *User Modeling and User-Adapted Interaction* 17 (3): 305-337.

Hawking, D., N. Craswell,P. Bailey, and K. Griffihs. 2001. Measuring search engine quality. *Information Retrieval* 4 (1): 33-59.

Hernon, P., and E. Altman. 1996. *Service quality in academic libraries*. Norwood, NJ: Ablex.

Huo, C., Y. Zhao, and W. Ren. 2017. User behavior sequence modeling to optimize ranking mechanism for e-commerce search. In *Proceedings of the 3rd International Conference on Communication and Information Processing:* 164-169. Tokyo, Japan.

Hughes, A. M. 1994. *Strategic database marketing*. Chicago: Probus Publishing Company.

Karimi, E., M. Babaee, and M. Hosseini Beheshti. 2018. Analysis of User query refinement behavior based on semantic features: user log analysis of Ganj database (IranDoc). *Human Information Interaction* 5 (3): 1-14.

Kathuria, A., B. J. Jansen, C. Hafernik, and A. Spink. 2010. Classifying the user intent of web queries using k-means clustering. *Internet Research*. *Internet Research*. 20 (5): 563-581.

Khanmohammadi, S., N. Adibeig, and S. Shanehbandy. 2017. An improved overlapping k-means clustering method for medical applications. *Expert Systems with Applications* 67:12-18.

Kotler, P. 1994. *Marketing management, analysis, planning, implementation, and control*. London: Prentice-Hall International.

Marcus, C. 1998. A practical yet meaningful approach to customer segmentation. *Journal of consumer marketing*. 15 (5): 494- .

Park, M., and T. S. Lee. 2013. Understanding science and technology information users through transaction log analysis. *Library hi tech*.

Park, S., J. H. Lee, and H. J. Bae. 2005. End user searching: A Web log analysis of NAVER, a Korean Web search engine. *Library and Information Science Research* 27 (2): 203-221.

Rabiei, M., S. M. Hosseini-Motlagh, and A. Haeri. 2017. Using text mining techniques for identifying research gaps and priorities: a case study of the environmental science in Iran. *Scientometrics* 110 (2): 815-842.

Rabiei, M., S. M. Hosseini-Motlagh, and B. Minaei Bidgoli. 2019. Using One-Class SVM for Scientific Documents Classification Case study: Iranian Environmental Thesis. *Iranian Journal of Information processing and Management* 34 (3): 1211-1234.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20: 53-65.

Stenmark, D. 2008. Identifying clusters of user behavior in intranet search engine log files. *Journal of the American Society for Information Science and Technology* 59 (14): 2232-2243.

Wang, Y., Q. Chen, C. Kang, and Q. Xia. 2016. Clustering of electricity consumption behavior dynamics toward big data applications. *IEEE transactions on smart grid* 7 (5): 2437-2447.

Wolfram, D. 2008. Search characteristics in different types of Web-based IR environments: Are they the same? *Information processing and management* 44 (3): 1279-1292.

Zhijing W., X. Xiaohui, L. Yiqun, Z. Min, & M. Shaoping. 2017. *A Study of User Image Search Behavior Based on Log Analysis.* In China Conference on Information Retrieval, pp. 69-80. Springer, Cham.

سمیه فتاحی

دانش‌آموختهٔ دکتری تخصصی دانشگاه تهران در رشتهٔ مهندسی کامپیوتر (گرایش هوش مصنوعی و رباتیک) است. ایشان هم‌اکنون استادیار پژوهشی گروه سیستم‌های اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران است.

مدل‌سازی کاربر، داده‌کاوی، تحلیل کلان‌داده، تشخیص الگو، یادگیری الکترونیکی و یادگیری ماشین از جمله علایق پژوهشی وی است.

محمد ربیعی

متولد سال ۱۳۶۲، دارای مدرک دکتری مهندسی فناوری اطلاعات از دانشگاه علم و صنعت ایران است. ایشان هم‌اکنون استادیار پژوهشکده فناوری اطلاعات، گروه پژوهشی کسب‌وکار الکترونیک پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

متن‌کاوی، پردازش زبان طبیعی، داده‌کاوی اطلاعات علم و فناوری و تحلیل رفتار کاربران از جمله علایق پژوهشی وی است.